BIT BUDGET: HOW MUCH INFORMATION DOES EEG ADD BEYOND A FROZEN LANGUAGE PRIOR?

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG-to-text systems can produce fluent language even when the underlying EEG signal is corrupted, raising a fundamental question: how much text-specific information does the neural signal contribute beyond a language prior? We introduce **Bit Budget**, a calibrated framework that treats neural-to-language decoding as a communication-channel measurement problem and quantifies evidence-dependent information beyond a frozen language prior. On the ZuCo corpus, gaze provides a significant information lower bound of **0.107 bits/token**, while the evaluated EEG representation yields **0.000 bits/token** with no significant evidence dependence. On the independent COFETT corpus, the same framework recovers an injected text-dependent signal of **0.149 bits/token**, whereas real EEG remains within ± 0.004 bits/token across the evaluated conditions. Notably, EEG still decodes four-way task/session identity with **82.1% balanced accuracy**, indicating that the representation contains structured information that does not translate into detectable token-level linguistic evidence. These findings expose a gap between language-generation performance and dependence on neural evidence, and motivate calibrated information measures and explicit evidence-destruction controls as complementary tools for evaluating neural-to-language systems.

1 INTRODUCTION

Electroencephalography (EEG) offers a non-invasive window into neural activity, yet language decoding from EEG remains notoriously difficult due to the signal's noisy, spatially coarse, and highly variable nature. In open-vocabulary EEG-to-text generation, this challenge is exacerbated by powerful language models (LMs) that can hallucinate fluent text even when the neural conditioning signal is uninformative. Recent literature demonstrates that generation performance often remains unchanged when EEG is corrupted or replaced by noise (Jo et al., 2024; 2025; Zhang et al., 2026b). Notably, on the ZuCo dataset (Hollenstein et al., 2018; 2020), Zhang et al. (2026b) report that real EEG performs indistinguishably from noise, whereas on the COFETT dataset, real EEG appears to outperform noise baselines. This discrepancy highlights a fundamental question that standard generation metrics fail to address: *how much information does the EEG signal actually contribute beyond the underlying linguistic prior?*

Meaningful evaluation requires isolating the information supplied by the biological measurement from the predictive power of the linguistic context. Estimating this added information from high-dimensional neural signals poses significant challenges: variational mutual information (MI) bounds are frequently loose and highly sensitive to critic architectures and optimization dynamics (Poole et al., 2019; Song & Ermon, 2020), while naive contrasts between matched and mismatched evidence do not admit a rigorous MI interpretation without strong assumptions regarding the negative distribution (Palomar & Verdú, 2008). These limitations necessitate an explicitly calibrated measurement protocol grounded by structured null controls.

To this end, we introduce **Bit Budget**, a framework for measuring the information contribution of recorded evidence to EEG-to-text decoding beyond a frozen language prior (instantiated via GPT-2 (Radford et al., 2019)). Our approach models an evidence-dependent residual that modulates token probabilities; the resulting log-probability ratio directly measures the signal's informative contribution. We estimate this quantity using a contrastive objective, rigorously test it against structured null distributions, and calibrate the estimator on synthetic channels of known capacity.

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.

Our empirical evaluation reveals three primary findings. First, behavioral gaze acts as a robust positive control, yielding an estimated 0.107 bits/token, whereas the evaluated fixation-locked EEG band-power representation yields 0.000 bits/token, non-significant against both structured nulls. We emphasize that this indicates an absence of *detectable* information under the evaluated setting, rather than proving EEG fundamentally lacks linguistic signal. Second, a replication of a prominent EEG-to-text system (Wang & Ji, 2022) produces essentially identical generation outputs regardless of whether the conditioning EEG is real, shuffled, or pure noise. Evaluation on COFETT follows the same pattern: real EEG supplies no reliable text information even under its own evaluation scheme, whereas an injected synthetic control signal is readily recovered. Finally, using likelihood-ratio candidate selection, gaze data achieves an information transfer rate of 11.1 bits/min at $N = 4$ and an Area Under the Risk-Coverage curve (AURC) of 0.056 (compared to 0.226 for a length/prior baseline). Because our pre-registered candidate-pool validity check failed at these small pool sizes, we report this specific result under a transparent logged override rather than as an unqualified confirmatory success.

In summary, our primary contributions are:

- **A calibrated framework for measuring evidence dependence.** We formulate EEG-to-text evaluation as a rigorous measurement problem relative to a frozen language prior. Our prior-residual critic employs an exact partition function and an explicit zero-information anchor, utilizing synthetic channels of known capacity to characterize empirical sensitivity.
- **A controlled analysis of EEG and gaze evidence.** We apply this framework to simultaneously recorded gaze and EEG. Gaze serves as a positive control (0.107 bits/token), while the evaluated EEG representation yields 0.000 bits/token under our experimental constraints. We further test these findings by comparing real, shuffled, and noise EEG in a reproduced generation pipeline, and on two independent corpora: the EEG result on COFETT and the gaze result on OneStop.
- **A communication-based evaluation of the measured signal.** We leverage the evidence-dependent likelihood ratio for candidate selection, quantifying the resulting communication rate and risk–coverage trade-off. By explicitly reporting the failure of the candidate-pool validity check alongside our logged override, we transparently bound the limitations of this evaluation.

2 Related Work

EEG-to-text research has moved from open-vocabulary sequence decoding (Wang & Ji, 2022) toward more structured neural–language representations, including discrete neural tokenization (Duan et al., 2023), contrastive alignment (Zhou et al., 2024), masked representation learning (Wang et al., 2024), diffusion-based decoding (Jeon & Kim, 2025), and semantic representations coupled with strong language priors (Liu et al., 2025; Yang et al., 2026). Related non-invasive language-decoding studies have also demonstrated semantic reconstruction from MEG and fMRI (Défossez et al., 2023; Tang et al., 2023). These advances improve linguistic reconstruction, but fluent output alone does not reveal how much a prediction depends on the recorded neural signal.

Recent work has therefore questioned whether neural decoding performance reflects genuine signal dependence or structure supplied by the decoder and its priors. EEG-to-text systems can retain similar performance when the EEG is corrupted, shuffled, or replaced by noise (Jo et al., 2024; 2025), while broader analyses have highlighted spurious reconstruction and source-attribution issues in brain-to-language decoding (Shirakawa et al., 2025; Zhang et al., 2026a). Zhang et al. (2026b) report no significant real-EEG versus noise difference on ZuCo, but observe a gap on the repeated-reading COFETT corpus. Simultaneous eye-tracking provides a useful complementary control because gaze measures reflect lexical and reading behavior (Dimigen et al., 2011; Hollenstein et al., 2022). These findings motivate evaluations that measure evidence dependence directly rather than infer it from generation quality.

Information-theoretic and contrastive estimators provide a principled route to quantifying evidence–target dependence (van den Oord et al., 2018; Belghazi et al., 2018; Poole et al., 2019), but their values can depend strongly on the critic, optimization procedure, and negative-sampling distribution (Song & Ermon, 2020). Bit Budget instead treats this as a measurement problem: evidence is evalu-

ated relative to a frozen language prior, with an exact zero-information anchor, structured nulls, and synthetic channels of known capacity. The resulting estimate is therefore interpreted conservatively as detectable evidence-dependent information under the specified estimator and null construction.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

We formulate neural-to-language decoding as a communication-channel measurement problem. Let W_t denote the target token at position t, C_t its preceding linguistic context, and $p_0(w_t \mid C_t)$ the frozen language-prior distribution for the realized token w_t. Let e_t denote the recorded evidence at position t (EEG, gaze, or both), and define the local evidence window as $E_t = (e_{t-1}, e_t, e_{t+1})$ (The gaze on the next word partly reflects the current word (spillover), and the current word is partly seen before it is fixated (parafoveal preview). So the model sees evidence for tokens $t-1$, t, and $t+1$). Our objective is to quantify the additional token-specific information supplied by E_t beyond what is already predictable from C_t, conceptually represented by

$$I(W_t; E_t \mid C_t). \tag{1}$$

We refer to the incremental text-specific information that can be certified from an evidence channel beyond the frozen prior as its *bit budget*. We further ask whether this information can support controlled discrimination of the correct text among competing alternatives.

3.2 FRAMEWORK ARCHITECTURE

Figure 1 summarizes the framework; we describe its components below.

Evidence encoder. The encoder f_ϕ maps the high-dimensional evidence window (of input dimension d_{in}) to a compact latent representation $u_t = f_\phi(E_t) \in \mathbb{R}^{d_u}$, where $d_u \ll d_{\text{in}}$. Its norm is constrained by $\|u_t\|_2 \le s_{\max}$, where $s_{\max}$ is the prescribed latent-norm bound, and the representation is centered using its running mean. For EEG, raw inputs are reshaped into frequency bands over channels, followed by band-specific spatial projection, attention pooling, and a temporal transformer over the evidence window. The fixation indicator is never supplied to the EEG encoder, preventing the model from exploiting the missingness pattern as a behavioral shortcut. For gaze, the behavioral feature vector is mapped by a linear layer (with layer normalization) into a temporal Transformer of the same architecture but with its own weights. For EEG+gaze, each modality is encoded separately, and the two latents are concatenated and linearly fused into u_t.

Prior-residual critic. Let V denote the vocabulary and $v \in V$ a candidate token. Given the latent evidence representation $u_t \in \mathbb{R}^{d_u}$, the critic modifies the frozen GPT-2 prior through an exponential tilt:

$$p_\phi(v \mid E_t, C_t) = \frac{p_0(v \mid C_t)\exp\{\gamma_t \ell_t(v)\}}{Z_t}, \qquad \ell_t(v) = \langle \Omega_v, Bu_t\rangle + \langle \varphi(\Omega_v), Au_t\rangle, \tag{2}$$

where p_ϕ is the evidence-conditioned distribution parameterized by the trainable residual components ϕ, $\langle\cdot,\cdot\rangle$ is the Euclidean inner product, Ω_v is the frozen GPT-2 output embedding of token v, B and A are trainable projection matrices matching the latent dimension d_u, $\varphi(\cdot)$ is a nonlinear MLP, and $\ell_t(v)$ is the evidence-dependent residual score. The exact partition function is

$$Z_t = \sum_{v \in V} p_0(v \mid C_t)\exp\{\gamma_t \ell_t(v)\}. \tag{3}$$

Computing Z_t over the full vocabulary avoids the truncation bias of top-k normalization.

The non-negative gain $\gamma_t = \gamma\, \mathbb{I}[e_t \text{ observed}]$, where $\mathbb{I}[\cdot]$ denotes the indicator function, controls the strength of the evidence correction. We use $\gamma = 1$ during training and the fixed evaluation gain $\gamma_{\text{eval}} = \text{softplus}(-2) \approx 0.127$, where $\text{softplus}(x) = \log(1 + e^x)$. The resulting per-token log-likelihood ratio is

$$T_t = \log \frac{p_\phi(w_t \mid E_t, C_t)}{p_0(w_t \mid C_t)} = \gamma_t \ell_t(w_t) - \log Z_t \le -\log p_0(w_t \mid C_t). \tag{4}$$

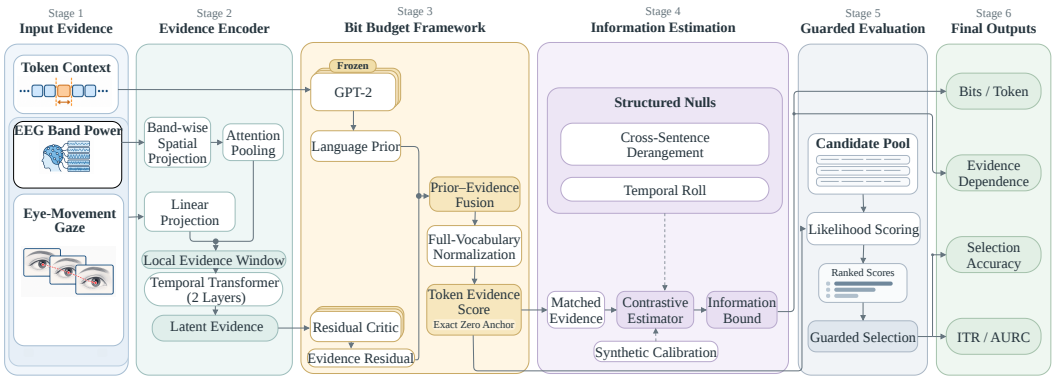


Figure 1: **Overview of the Bit Budget framework. Stage 1:** token context feeds the frozen GPT-2 prior (Qwen2.5-0.5B for Chinese COFETT), while EEG band power and gaze form the recorded evidence. **Stage 2:** modality-specific projections and a two-layer temporal Transformer over the ± 1-token window (separate weights per modality) produce the bounded latent u_t. **Stage 3:** a prior-residual critic exponentially tilts the prior with full-vocabulary normalization, giving a token evidence score with $T_t = 0$ when evidence is ignored. **Stage 4:** structured nulls (derangement, temporal roll) and synthetic calibration support a conservative contrastive information bound. **Stages 5–6:** candidate sentences are scored for guarded selection, yielding bits/token, evidence dependence, selection accuracy, and ITR/AURC.

When evidence is ignored by setting $\gamma = 0$, $T_t \equiv 0$, providing an exact zero anchor. We write $T_t(v, E)$ for equation 4 evaluated at token v with evidence window E, and abbreviate $T_t(E) = T_t(w_t, E)$.

Estimator and null distributions. During training, each matched token–evidence pair is contrasted against $M_{\text{tr}} = 4$ in-batch deranged evidence samples. At evaluation, we use $M = 1{,}000$ structured null draws. Let $E_t^{(m)}$ denote the mth null evidence sample for position t, with $m \in \{1, \ldots, M\}$. The contrastive information-bound estimate is

$$\hat{I} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left[T_t(E_t) - \log \left(\frac{e^{T_t(E_t)} + \sum_{m=1}^{M} e^{T_t(E_t^{(m)})}}{M + 1} \right) \right], \tag{5}$$

where $\mathcal{T}$ is the set of evaluated token positions and $\hat{I}$ is computed in nats (natural-log units) and divided by $\ln 2$ to report bits/token. We also report the Donsker–Varadhan (DV) estimate $\hat{I}_{\text{DV}} = \frac{1}{|\mathcal{T}|} \sum_t T_t(E_t) - \log \frac{1}{M|\mathcal{T}|} \sum_{t,m} e^{T_t(E_t^{(m)})}$ (Donsker & Varadhan, 1983; Belghazi et al., 2018).

We use equation 5 instead of the uncalibrated statistic $\mathbb{E}[T_t(E_t)] - \mathbb{E}[T_t(E_t^{(m)})]$, whose expectation can contain an additional lautum-information term (Palomar & Verdú, 2008) and is sensitive to score scaling. Under the standard InfoNCE construction, marginally sampled negatives admit the usual variational lower-bound interpretation. Our evaluation uses two structured nulls: *cross-sentence derangement*, which breaks text-specific correspondence while preserving coarse structure, and a *temporal roll*, which preserves subject/session characteristics while disrupting token alignment. Because these nulls are deliberately structured, we report $\hat{I}$ as a *calibrated contrastive information-bound estimate under the specified null construction, rather than as an unrestricted estimate of* $I(W_t; E_t | C_t)$. A channel is considered informative only when the estimate is positive and significant against both nulls (pre-registered hypothesis H1).

Calibration before measurement. Before biological-signal analysis, we evaluate the estimator on synthetic channels with known capacity b bits/token, where b denotes the injected channel capacity. Synthetic sentences are sampled from p_0 and paired with Gaussian evidence codes of controlled capacity. The pre-biological calibration verifies the zero anchor and lower-bound soundness. A subsequent low-capacity sweep uses independently trained seeds to characterize sensitivity at small b without retuning the biological-channel models.

Likelihood-ratio selector. Let $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(N)}\}$ denote a candidate pool of N sentences, with $y = (y_1, \ldots, y_{|y|})$ denoting one candidate. Its likelihood-ratio score is

$$S(y) = \sum_{t=1}^{|y|} T_t(y_t, E_t), \qquad \hat{y} = \arg\max_{y \in \mathcal{Y}} S(y), \tag{6}$$

where $S(y)$ is the sentence-level evidence score and $\hat{y}$ is the selected candidate. Distractors are matched in sentence length and prior probability. We denote by N^* the candidate-pool size frozen from the validation-stage E2 measurement before test access. Because the text-only pool-validity test subsequently fails at small pool sizes, the E4 results are reported under the corresponding logged override (more details in §4.4).

4 Experimental Setup

We evaluate recorded evidence on a primary corpus (ZuCo) and two external corpora, COFETT (repeated-reading EEG) and OneStop (eye tracking), using controls designed to expose statistical shortcuts and prior-only behavior. All three corpora are summarized in Appendix A.

4.1 Datasets and Feature Extraction

ZuCo (primary). We use ZuCo 1.0 and 2.0 (Hollenstein et al., 2018; 2020), which provide synchronized EEG and eye-tracking recordings during natural English sentence reading. The corpus contains 18,890 readings of 1,319 unique sentences from 30 participants, with no within-participant repetitions (Zhang et al., 2026b). For each word, we extract an 840-dimensional EEG representation comprising band power from 105 scalp channels across eight frequency bands, and a 6-dimensional gaze representation containing fixation and standard reading-time measures. Unfixated words, which account for about 37% of words, are masked rather than zero-imputed, and GPT-2 sub-tokens inherit the evidence and mask of their parent word. See Appendix A.

COFETT (external EEG). We additionally evaluate COFETT (Zhang et al., 2026b) (OpenNeuro ds006317), a Chinese EEG-only corpus designed to address repeated-reading effects. It contains 1,124 unique sentences read by two participants, including a 252-sentence subset repeated across four sessions. We extract the same 840-dimensional band-power representation from each character's 0.4 s window; recall windows and preprocessing are described in Appendix B.

OneStop (external gaze). To assess gaze generalization beyond ZuCo, we apply the unchanged instrument to OneStop (Berzak et al., 2025), an English eye-tracking corpus containing 9,718 readings of 126 newswire texts from 180 readers. We retain the same GPT-2 Large prior and 6-dimensional gaze representation and use text-disjoint train/validation/test splits of 7,618/960/1,140 readings. Further details are provided in Appendix C.

4.2 Text-Disjoint Splits and Leakage Control

Each distinct sentence (a *text*) is read by several participants (*readings*), so a random split would place the same text in both training and test. We therefore assign all readings of a text to one partition, verified programmatically for every build. ZuCo is split into 15,131 / 1,834 / 1,925 readings (133 test texts); validation texts are further halved into a checkpoint-selection half and a measurement half (970 readings of 68 texts) used for all E2 estimates. In an instance-split check, where test texts may appear in training, gaze gives a similar bound (0.099 vs. 0.107 bits/token) and EEG still shows no signal (0.004 bits/token with a negative DV bound, no selection signal). COFETT is evaluated under two splits: *unseen sentences*, a text-disjoint split of session 1 (3,382 / 424 / 458 readings), and COFETT's own scheme, in which test items are repetitions 15–16 of sentences read in training (validation: 13–14). As in COFETT, we train one model per participant, and additionally a pooled model.

4.3 Model Architecture: Frozen Prior and Evidence Encoder

The frozen prior p_0 is GPT-2 Large (Radford et al., 2019) for ZuCo and Qwen2.5-0.5B (Qwen Team, 2024) for Chinese COFETT; only its vocabulary and embedding sizes change the size of the tilt (Appendix B). Its parameters are never updated, so it is a fixed baseline against which evidence-conditioned deviations are measured. The encoder f_ϕ processes the evidence window $(t-1, t, t+1)$: EEG bands undergo band-wise spatial projection and band-attention pooling, gaze is linearly projected, and a two-layer temporal Transformer (Vaswani et al., 2017) (same architecture per modality, separate weights) produces the latent u_t that parameterizes the tilt; for EEG+gaze, the two modality latents are concatenated and linearly fused. All other components and hyperparameters are identical across corpora.

4.4 Gated Evaluation Pipeline

We use a sequential, logged evaluation pipeline: **E0** constructs text-disjoint splits and verifies zero text overlap; **E1** calibrates estimator soundness, the zero anchor, and sensitivity on synthetic channels of known capacity (2,000 training sentences per capacity, 20% of tokens hidden); **E2** measures calibrated contrastive information bounds and permutation p-values on held-out data; **E3** determines and freezes the channel-specific candidate-pool size before test access ($N^* = 4$ for gaze/EEG+gaze and $N^* = 2$ for EEG); and **E4** evaluates the frozen models on the test set. Since the pre-specified pool-validity criteria are not satisfied at these candidate sizes, E4 is reported under a logged override and interpreted together with the corresponding control arms. COFETT is evaluated with the same E1-calibrated instrument and E2–E4 procedures. Training details, evaluation metrics, and hyperparameters are provided in Appendix I.

5 Results and Discussion

We follow the gated evaluation pipeline (E1–E4), progressing from instrument calibration to biological evidence measurement and controlled communication evaluation. This ordering separates estimator validity from downstream decoding performance.

5.1 Instrument Calibration (E1)

We first evaluate the estimator on synthetic channels with known information capacity. To probe the low-information regime relevant to our biological measurements, we sweep the injected capacity over $b \in \{0, 0.005, 0.01, 0.025, 0.05, 0.10, 0.15, 0.25, 0.5, 1, 2\}$ bits/token using three independent training seeds. At $b = 0$, the recovered estimate is -0.0022 ± 0.0010 bits/token, and none of the three seeds passes both null tests. Across the sweep, the estimate never exceeds the injected capacity. Detection is limited at low capacities: $b = 0.01$–0.15 is detected in none of the three seeds, while $b = 0.25$ is detected in two of three seeds. In contrast, all three seeds pass both null tests for $b \in \{0.5, 1, 2\}$ bits/token. Per-capacity values are given in Appendix D.

These results establish the expected zero behavior and show that the estimator is conservative, but has limited sensitivity to very small information rates. Given the small number of seeds and non-monotonic detection at low capacities, we do not interpret the sweep as a formal detection threshold. Instead, it provides an empirical sensitivity characterization for the estimator used in our biological experiments. Thus, a near-zero biological estimate should be interpreted as *no detectable information under the evaluated representation and estimator*, rather than as evidence that the underlying mutual information is exactly zero. Figure 2 summarizes information recovery and detection across the injected capacities.

5.2 Evidence Dependence: Gaze vs. EEG (E2)

We evaluate the held-out ZuCo measurement half (970 readings of 68 texts). Table 1 and Figure 3 show a clear behavioral–neural contrast. The 6-dimensional gaze record yields a significant **0.107 bits/token** lower bound (95% CI: [0.090, 0.123]), passing both nulls ($p = 0.001$ for derangement and temporal roll), with positive estimates across all 30 readers and four ZuCo tasks. A structure-only control that preserves sentence length, word grouping, and sub-token replication but random-

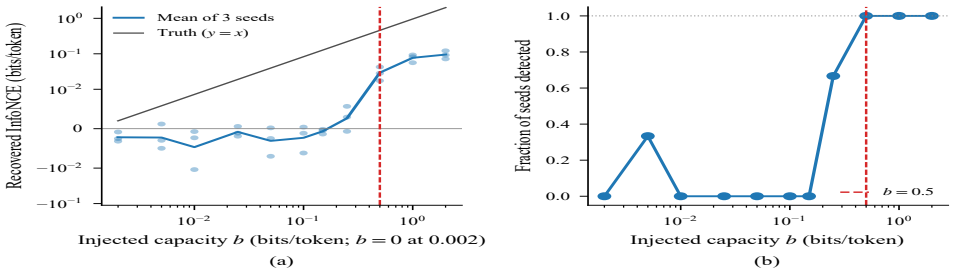


Figure 2: Synthetic-channel calibration and low-capacity sensitivity. No recovered synthetic estimate exceeds the injected capacity, while detection is inconsistent below $b = 0.5$ bits/token in the three-seed follow-up sweep.

Table 1: E2 results on the held-out measurement half (InfoNCE: calibrated information bound, bits/token). Fano bits are from the E4 test (seed 0).

CHANNEL	InfoNCE	DV	p (DER./ROLL)	TOKENS/SENT	N^*	FANO BITS	H1
Gaze (6-d)	0.107	0.080	0.001 / 0.001	24.6	4	1.03	PASS
EEG (840-d)	0.000	−0.000	0.270 / 0.254	16.0	2	0.00	FAIL
EEG + Gaze	0.102	0.079	0.001 / 0.001	24.6	4	1.00	PASS

izes gaze content yields 0.011 bits/token. This suggests that segmentation structure alone does not explain the gaze effect, although the difference between two lower bounds is not treated as an exact decomposition.

In contrast, the 840-dimensional fixation-locked EEG band-power representation yields **0.000 bits/token**, with non-significant derangement and temporal-roll tests ($p = 0.270$ and $p = 0.254$). We therefore find no detectable token-level information for this representation under the present estimator. Because the low-capacity calibration (Section 5.1) shows that weak channels can remain undetected, this near-zero value is not evidence of exactly zero underlying EEG mutual information. The result is consistent with Zhang et al. (2026b), remains stable across alternative EEG windows and alignments, and coexists with detectable stronger spike-in signals (Appendix G.5). EEG+gaze yields 0.102 bits/token versus 0.107 for gaze alone; thus, the evaluated EEG representation does not increase the reported bound, without implying negative or zero conditional EEG information. On OneStop, gaze again passes both nulls with 0.114 bits/token (95% CI [0.100, 0.127]), but a structure-only control already reaches 0.091 (vs. 0.011 on ZuCo), so most of the OneStop bound reflects text layout rather than gaze content (Appendix C).

5.3 Selection Accuracy and Pool Validity (E3)

We assess sentence selection performance across varying candidate pool sizes (N). As anticipated, gaze-based selection accuracy degrades as pool size increases—dropping from 92.4% at $N = 2$ to 34.4% at $N = 64$—yet it consistently outperforms random chance. Conversely, EEG-based selection fails to exceed chance-level performance across all evaluated values of N.

Our pre-registered pool-validity criterion mandates that the strongest text-only heuristic (e.g., sentence length, type-token ratio, or prior probability) must remain within two percentage points of random chance. However, this validity gate fails on the test pools used in our final evaluation (E4), where the text-only baselines achieve 30.8% at $N = 4$ (chance 25%) and 56.4% at $N = 2$ (chance 50%). Consequently, while we proceed with the pre-determined frozen pool sizes ($N^* = 4$ for gaze and EEG+gaze, and $N^* = 2$ for EEG), we report the results of E4 under a logged override. We note that the larger validation pools ($N \in \{32, 64\}$) successfully satisfy the strict validity criterion. Notably, real EEG evidence never surpasses the performance of its zeroed-out ablation, whereas gaze remains robustly above chance at all scales (e.g., 34.4% vs. 1.6% at $N = 64$). Comprehensive results for this stage are provided in Appendix E.

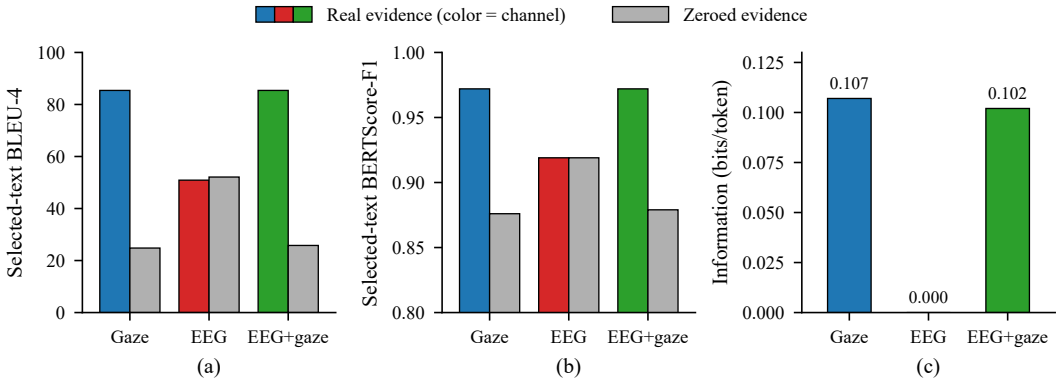


Figure 3: **Language quality versus evidence dependence.** Bar color denotes the evidence channel (blue: gaze; red: EEG; green: EEG+gaze); colored bars use real evidence and grey bars use zeroed evidence. **(a)** BLEU-4 and **(b)** BERTScore-F1 of the candidate selected on the test pools (E4; $N^* = 4$ for gaze and EEG+gaze, $N^* = 2$ for EEG); these are functions of selection accuracy, not generation metrics. The y-axis in (b) starts at 0.80. **(c)** Calibrated InfoNCE information-bound estimate on the validation measurement half (E2). EEG text similarity is unchanged when its evidence is zeroed, and its information estimate is 0.000 bits/token, whereas gaze drops sharply without evidence and carries 0.107 bits/token.

Table 2: E4 test-set results under the logged pool-validity override. Gaze accuracy and T_2 ITR are three-seed means; other entries are seed 0.

QUANTITY	GAZE	EEG	EEG+GAZE
Frozen N^*	4	2	4
Accuracy	0.832 ± 0.016	0.496	0.812
Fano bits	1.03	0.00	1.00
ITR (T_2), bits/min	11.1 ± 0.6	0.0	10.3
ITR, T_1 / T_3	16.2 / 14.6	0.0 / 0.0	15.7 / 14.2
AURC / baseline	0.056 / 0.226	0.489 / 0.440	0.061 / 0.230
Confidence AUC / pAUC	0.841 / 0.680	0.531 / 0.488	0.837 / 0.692
BLEU-4, real / zeroed[†]	85.4 / 24.8	50.9 / 52.1	85.4 / 25.8
H2: matched-N	FAIL	FAIL	FAIL
H3: ITR 1–60 bits/min	PASS	FAIL	PASS
H4: pAUC > 0.5	PASS	FAIL	PASS
H5: TOST $\pm$2pp	Underpowered	Underpowered	Underpowered

[†]Computed on the selected candidate; not a generation metric.

5.4 Logged Test Evaluation (E4)

Table 2 summarizes the test evaluation at the frozen channel-specific pool sizes. Because the pool-validity criterion fails at these small N, the results are treated as controlled diagnostic evidence rather than an unqualified confirmatory test. At $N = 4$, gaze achieves $\mathbf{83.2 \pm 1.6}\%$ accuracy across three seeds, while its six no-evidence controls remain substantially lower. It also yields $\mathbf{11.1 \pm 0.6}$ bits/min ITR and an AURC of **0.056**, compared with 0.226 for the length/prior baseline. Because the pool-validity gate fails and TOST is underpowered, these controls are not interpreted as formal equivalence to chance. At $N = 2$, EEG achieves 49.6% accuracy, identical to its zeroed-evidence arm (49.6%) and close to the remaining controls, providing no observed evidence dependence in this evaluation. The six gaze control arms range from 0.226 to 0.323 (chance 0.250); the full control battery and communication-system plots are given in Appendix F.

5.5 External Check on a Repeated-Reading Corpus (COFETT)

On COFETT, Zhang et al. (2026b) report better EEG-to-text performance with real than noise EEG. We apply the unchanged instrument (§ 4) to fourteen conditions: reading and recall for each participant (S1, S2) and a pooled model under both evaluation splits ($2 \times 3 \times 2 = 12$), plus two pooled conditions using a smaller session-1 version of COFETT's scheme. The instrument remains sensitive on this corpus. A text-dependent signal injected at one tenth of the EEG feature noise ($\alpha = 0.1$) is recovered at 0.149 bits/token, increasing to 0.327 and 0.703 bits/token for stronger injections. Real EEG, however, remains within ± 0.004 bits/token across all fourteen conditions (Appendix B), with H1 failing in 11, including 5/6 conditions under COFETT's scheme and 5/6 on unseen sentences. Recall fails in all seven conditions, while unseen-sentence selection remains at or near chance for $N \in \{4, 8, 16, 32\}$.

The three H1-passing cells occur only during reading and yield $+0.001$–$+0.003$ bits/token, at least $50\times$ smaller than the injected control. They are uncorrected for multiple comparisons and do not replicate across participants or evaluation designs; the pooled 16-repetition model fails in both phases. The null finding also persists across evidence-window choices, tilt capacities from linear to rank 64, and evaluation gains, as well as under leave-one-subject-out evaluation, where both held-out participants remain at chance. A separate one-shot test session similarly gives 0.513 accuracy for reading and 0.504 for recall at $N^* = 2$ (chance $= 0.500$). A real-versus-noise gap in an embedding-similarity metric can reflect participant, session, or repetition structure rather than text-specific evidence. Our structured nulls preserve these factors while testing information beyond the language prior. Under the evaluated band-power representation, we therefore find no reliable text-dependent information in COFETT EEG as well.

5.6 Additional Analyses

Task/session probing. A four-way linear probe on sentence-level EEG distinguishes ZuCo reading paradigms with **0.821 balanced accuracy** (chance: 0.25). Accuracy remains significant at 0.482 ($p = 0.005$) after residualizing gaze covariates, indicating EEG captures macro-level session artifacts despite lacking detectable token-level linguistic signal.

Gaze scalability and LOSO. In a separate retraining run, gaze accuracy at $N = 64$ rises (18.3% $\rightarrow$ 30.5% $\rightarrow$ 42.0%) with 25%, 50%, and 100% of training data, showing no saturation. Leave-one-subject-out (LOSO) evaluation across 12 ZuCo 1.0 readers yields 0.792 ± 0.067 accuracy at $N = 4$ and 0.466 ± 0.084 at $N = 32$, far exceeding zeroed baselines (0.256 and 0.025; Appendix G.6).

Baseline reproduction. We replicate an open-vocabulary EEG-to-text pipeline (Wang & Ji, 2022). Text outputs are completely invariant to whether the conditioning EEG is real, temporally shuffled, or pure noise (e.g., BLEU-1 is identically 13.04 and 19.17 across arms; Appendix H), confirming negligible dependence on the biological input.

6 Conclusion

We introduced **Bit Budget**, a calibrated framework for measuring text-specific information contributed by an evidence channel beyond a frozen language prior. Gaze provides a positive control, yielding 0.107 bits/token and supporting evidence-dependent candidate selection, whereas the evaluated fixation-locked EEG representation yields 0.000 bits/token with no significant evidence dependence. The EEG null replicates on the independent COFETT corpus, where injected control signals are readily recovered, and the gaze bound replicates on OneStop. These results indicate no detectable token-level linguistic information in the tested EEG representation under our estimator and evaluation protocol, rather than establishing that EEG contains no such information. More broadly, our results highlight the need to evaluate neural-to-language systems through calibrated information bounds and explicit evidence-destruction controls, alongside conventional generation metrics.

GENERATIVE AI DISCLOSURE

Generative AI tools were used in this work only to assist with language polishing of the manuscript. All scientific content, methodology, experiments, results, interpretation, and final verification were carried out by the authors. The authors reviewed all AI-assisted text and figure content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce the experiments throughout the main paper and appendix. Section 4 describes the ZuCo datasets, preprocessing, text-clustered splits, feature construction, training protocol, and gated evaluation procedure. Complete hyperparameters are reported in Appendix I.3, while Appendices D–G provide the synthetic calibration, candidate-selection results, test controls, and additional diagnostic analyses. Appendix H documents the reproduced EEG-to-text baseline, and Appendix J records the test-split access protocol and experiment provenance. All reported experiments use fixed random seeds and logged configurations, and the low-capacity calibration is identified separately as a post-campaign sensitivity analysis that does not retune the biological-channel models. Code, configuration files, and run artifacts needed to reproduce the reported results — including the driver scripts for the main campaign and for the leave-one-subject-out, per-subgroup, instance-split, structure-only-control, attribute-probe, and baseline-generation analyses — are included in the supplementary material. The COFETT analysis uses public OpenNeuro data (ds006317, CC0) and the same code base; only the data adapter and prior differ (Appendix B). Anonymized source code is available at `https://anonymous.4open.science/r/eeg_check-1E1E` for the duration of review.

REFERENCES

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540, 2018.

Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. OneStop: A 360-participant English eye tracking dataset with different reading regimes. *Scientific Data*, 12 (1):1995, 2025. doi: 10.1038/s41597-025-06272-2. URL `https://doi.org/10.1038/s41597-025-06272-2`.

Xiaogang Chen, Yijun Wang, Masaki Nakanishi, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao. High-speed spelling with a noninvasive brain–computer interface. *Proceedings of the National Academy of Sciences*, 112(44):E6058–E6067, 2015. doi: 10.1073/pnas.1508080112.

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.

Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M. Jacobs, and Reinhold Kliegl. Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General*, 140(4):552–572, 2011.

Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. IV. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

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

Nora Hollenstein, Jän 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(1):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 (LREC)*, pp. 138–146, Marseille, France, 2020. European Language Resources Association (ELRA).

Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A. Jäger, and Nicolas Langer. The ZuCo benchmark on cross-subject reading task classification with EEG and eye-tracking data. *Frontiers in Psychology*, 13:1028824, 2022.

Mingyu Jeon and Hyobin Kim. DELTA: Language diffusion-based EEG-to-text architecture. *arXiv preprint arXiv:2511.21746*, 2025.

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, 2025. doi: 10.1038/s41598-025-29587-x.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Thomas E. Nichols and Andrew P. Holmes. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25, 2002. doi: 10.1002/hbm.1058.

Daniel P. Palomar and Sergio Verdú. Lautum information. *IEEE Transactions on Information Theory*, 54(3):964–975, 2008.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 5171–5180, 2019.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Ken Shirakawa, Yoshihiro Nagano, Misato Tanaka, Shuntaro C. Aoki, Yusuke Muraki, Kei Majima, and Yukiyasu Kamitani. Spurious reconstruction from brain activity. *Neural Networks*, 190:107515, 2025. doi: 10.1016/j.neunet.2025.107515.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26:858–866, 2023. doi: 10.1038/s41593-023-01304-9.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing EEG-to-text decoding through transferable representations from pre-trained contrastive EEG-text masked autoencoder. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7278–7292, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.393.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022. doi: 10.1609/aaai.v36i5.20472.

Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

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.

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

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, San Diego, California, United States, 2026b. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.61. URL https://aclanthology.org/2026.acl-long.61/.

Jinzhao Zhou, Yiqun Duan, Fred Chang, Thomas Do, Yu-Kai Wang, and Chin-Teng Lin. BELT-2: Bootstrapping EEG-to-language representation alignment for multi-task brain decoding. *arXiv preprint arXiv:2409.00121*, 2024.

A Datasets and Cross-Dataset Summary

Table 3 summarizes the three corpora, while Table 4 reports the main results obtained with the same instrument across datasets. The encoder, prior-residual tilt, estimator, nulls, gates, and training settings remain unchanged; only the data adapter and, for the Chinese corpus, the frozen language prior differ.

ZuCo features. For each fixated word, EEG band power is computed over 105 scalp channels in eight bands, $\theta_{1,2}$ (4–8 Hz), $\alpha_{1,2}$ (8.5–13 Hz), $\beta_{1,2}$ (13.5–30 Hz), and $\gamma_{1,2}$ (30.5–49.5 Hz), giving an 840-dimensional vector. The 6-dimensional gaze vector is $[\mathbb{I}_{\text{fixated}}, \log(1 + n_{\text{Fix}}), \log(1 + \text{FFD}), \log(1 + \text{GD}), \log(1 + \text{TRT}), \log(1 + \text{GoPT})]$, i.e., fixation indicator, fixation count, and first-fixation, gaze, total-reading, and go-past durations. Unfixated words (about 37% of words) have no fixation-locked EEG; their EEG rows are masked rather than zero-imputed, and GPT-2 sub-tokens inherit the evidence row and mask of their parent word.

Table 3: The three corpora used in this paper.

ITEM	ZUCO (PRIMARY)	COFETT (EXTERNAL)	ONESTOP (EXTERNAL)
Language	English	Chinese	English
Recorded signal	EEG + eye tracking	EEG only	eye tracking only
Participants	30	2	180
Unique texts	1,319 sentences	1,124 sentences	126 paragraph texts
Readings	18,890	4,264 (session 1); 10,312 (16 repetitions)	9,718
Reading task	natural sentence reading	one character every 0.4 s, then silent recall	natural paragraph reading
Evidence unit	word (fixation-locked)	character (0.4 s window)	word
Features	840-d EEG band power; 6-d gaze	840-d EEG band power	6-d gaze
Frozen prior	GPT-2 Large	Qwen2.5-0.5B (float32)	GPT-2 Large
Split (train / val / test readings)	15,131 / 1,834 / 1,925	unseen: 3,382 / 424 / 458; COFETT scheme: repetitions 1–12 / 13–14 / 15–16	7,618 / 960 / 1,140
Source	Hollenstein et al. (2018; 2020)	OpenNeuro ds006317, CC0 (Zhang et al., 2026b)	OSF, CC-BY (Berzak et al., 2025)

Table 4: The same instrument on three corpora (held-out measurement data). Bits/token beyond the prior under the derangement null. "Structure-only" keeps sentence length and word grouping but scrambles gaze content. †Pooled reading phase, unseen sentences; 11 of 14 COFETT conditions fail H1 (Appendix B).

DATASET	SIGNAL	BITS/TOKEN [95% CI]	p (DER./ROLL)	STRUCTURE-ONLY	H1
ZuCo	gaze	0.107 [0.090, 0.123]	0.001 / 0.001	0.011	PASS
ZuCo	EEG	0.000	0.270 / 0.254	–	FAIL
OneStop	gaze	0.114 [0.100, 0.127]	0.001 / 0.001	0.091	PASS
COFETT†	EEG	−0.003	0.789 / 0.001	–	FAIL
COFETT	injected control, $\alpha = 0.1$	0.149	0.003 / 0.003	–	PASS

B External Check on COFETT

Data and preprocessing. This section complements Section 4.1. COFETT provides continuous 128-channel EEG at 1 kHz. We match trials to sentences using the released presentation order and verify them by timing: the reading-to-recall interval equals $0.4\,\mathrm{s}$ times the number of characters, within 0.6 s; no trial is dropped. We notch-filter EEG at 50 Hz and resample to 250 Hz. Removing the 13-channel peripheral ring from the 124 EEG channels leaves 111; we then drop the last six (right-hemisphere) channels to obtain 105, matching ZuCo's 840-dimensional input. For each of the eight ZuCo bands, we compute the squared Hilbert envelope of the band-passed signal, average it within each character's 0.4 s window, and apply a logarithm. Recall timing is not recorded, so recall-phase character windows are spread uniformly over the recall interval; punctuation is masked like an unfixated word.

Prior and model sizes. We use Qwen2.5-0.5B in float32 as the frozen prior. As with GPT-2, we verify that it is causal, deterministic, and better than uniform prediction (mean surprisal 6.71 vs. 11.93 nats). The estimator, nulls, controls, gates, encoder, and training settings remain unchanged from ZuCo; only prior-dependent dimensions differ. Qwen uses an output vocabulary of 151,936 (151,665 tokenizer entries) and an embedding dimension of 896, compared with 50,257 and 1,280 for GPT-2, so φ changes from $1280 \to 64 \to 16$ to $896 \to 64 \to 16$. Main rows use 1,000 permutation draws and control rows 300.

Table 5: COFETT: bits/token beyond the prior, with p-values (derangement / roll). *H1 pass (both $p < 0.05$).

SPLIT	MODEL	RECALL (INNER SPEECH)	READING
COFETT scheme (16 reps; test = 15–16)	S1	-0.0014 (0.744 / 0.498)	$+0.0017$ (0.003 / 0.012)*
	S2	$+0.0000$ (0.022 / 0.469)	$+0.0005$ (0.054 / 0.001)
	Pooled	-0.0007 (0.550 / 1.000)	-0.0005 (0.323 / 0.001)
Session-1 repeated	Pooled	$+0.000$ (0.171 / 0.362)‡	$+0.001$ (0.017 / 0.001)*‡
Unseen sentences (text-disjoint)	S1	-0.0035 (0.926 / 0.947)	$+0.0003$ (0.160 / 0.534)
	S2	$+0.0000$ (0.311 / 0.236)	$+0.0028$ (0.038 / 0.008)*
	Pooled	-0.0000 (0.743 / 0.569)	-0.0026 (0.789 / 0.001)

‡Reported to three decimals.

Splits. We evaluate both split types described in Section 4.2. The unseen-sentence split uses session 1 from both participants (4,264 readings; 1,124 unique sentences). COFETT's repeated-reading scheme adds sessions 2–4, yielding 16 repetitions of a 252-sentence list (each list sentence also appears once among the 1,124 session-1 sentences), so every test sentence appears 13 times in training. We also evaluate a smaller session-1 repeated split with three earlier training presentations and one validation repetition (3,256/504/504 readings, pooled). Per-participant models use 2,132 readings for the unseen split and 5,156 for the 16-repetition split; the pooled 16-repetition set contains 10,312 readings.

Controls and selection. Table 5 lists all fourteen real-EEG cells; the injected controls, run on the reading phase, whose character timing is exact, are listed in Table 6. The injected control adds a known text-dependent signal to the reading-phase features, scaled by α relative to the feature noise. It is recovered at every strength, so the null on real EEG is not a failure of the pipeline. Because it is text-dependent, whereas the ZuCo spike-in of Appendix G.5 is gaze-derived, the two α scales are not comparable. On unseen sentences, selection accuracy is at or near chance ($N = 4$: 0.241 vs. 0.250; $N = 8$: 0.107 vs. 0.125; $N = 16$: 0.062 vs. 0.062; $N = 32$: 0.045 vs. 0.031), as is the zeroed-evidence control.

Robustness checks. The null does not depend on analysis choices (Table 6). Changing the evidence window or the tilt capacity leaves every estimate at or below zero. Larger evaluation gains only push the bound further negative, as expected when the tilt amplifies noise rather than signal. The smaller Qwen-sized φ does not hide information, since the injected control is recovered with it.

Leave-one-subject-out, data scaling, and one-shot test. Training on one participant and testing on the other gives chance-level selection for both held-out participants (Table 7), whereas ZuCo gaze reaches 0.792 at $N = 4$ under the same protocol (Appendix G.6). As the training set grows, unseen-sentence accuracy at $N = 32$ rises only from 1.8% to 3.6% to 5.4% (chance 3.1%). In a one-shot test session at the frozen pool size $N^* = 2$, pooled accuracy is 0.513 for reading and 0.504 for recall (chance 0.500).

Passing cells and multiple comparisons. Across the 14 real-EEG cells, H1 passes in three: unseen reading S2, 16-repetition reading S1 (DV $+0.0016$), and session-1 repeated reading. No multiple-comparison correction was applied; for the derangement null alone, about 0.7 passes would be expected by chance across 14 cells, so a tiny reading-phase effect cannot be excluded. However, every passing value lies between $+0.001$ and $+0.003$ bits/token, all are in the reading phase, none replicates across participants or designs, and the largest test (pooled, 16 repetitions) fails in both phases. Two of the three passes occur when test sentences were read before, which would fit recognition of repeated sentences, as the ablation of Zhang et al. (2026b) suggests; one occurs on unseen sentences, so we draw no conclusion about its source.

Table 6: COFETT robustness and injected controls, pooled reading phase, unseen sentences (bits/token beyond the prior). Non-main rows use 300 permutation draws; –: not reported.

FACTOR	SETTING	BITS/TOKEN	p (DER./ROLL)
Main	window ±1, rank 16, γ_{eval}	-0.0026	0.79 / 0.001
Window	current character only	-0.0000	–
	±3 characters	-0.0000	–
Tilt	linear (rank 0)	$+0.0000$	–
	rank 4	-0.0008	0.55 / 0.47
	rank 8	-0.0008	0.54 / 0.003
	rank 32	-0.0030	0.68 / 0.003
	rank 64, MLP width 256	-0.0000	0.56 / 0.57
Evaluation gain	$\gamma = 1$	-0.089	–
	$\gamma = 2$	-0.292	–
Injected control (text-dependent)	$\alpha = 0.1$	$+0.1491$	0.003 / 0.003
	$\alpha = 0.3$	$+0.3267$	0.003 / 0.003
	$\alpha = 1.0$	$+0.7030$	0.003 / 0.003

Table 7: COFETT leave-one-subject-out selection (reading phase).

HELD-OUT	N	ACCURACY	ZEROED	CHANCE
S1	4	0.232	0.214	0.250
S1	32	0.036	0.054	0.031
S2	4	0.250	0.250	0.250
S2	32	0.018	0.018	0.031
Mean	4	0.241	0.232	0.250
Mean	32	0.027	0.036	0.031

Roll-only significance. Some reading cells beat the roll null but not the derangement null. The roll null shifts a sentence's own evidence in time, whereas the derangement null uses another sentence's evidence with the same timing pattern. Beating only the roll null is therefore what a timing response shared by all sentences (e.g., to sentence onset) would produce, not text content, which would beat both. We did not isolate this component.

Scope and integrity. Two participants, band-power features per character, 105 of 124 channels, and a Chinese prior. We do not claim that COFETT EEG contains no information of any kind, nor that the COFETT score is wrong. An automated audit of stored channel lists found one run (participant 2, session 1, list 3) extracted with a wrong channel list after a failed metadata download (two EOG electrodes had been labelled EEG). It was re-extracted, and every affected dataset and experiment was rebuilt and rerun from scratch; participant-1 results were unaffected.

C External Check on OneStop (Gaze)

Data. OneStop (Berzak et al., 2025) is an English eye-tracking corpus of natural newswire reading without EEG. We use the ordinary-reading subset from 180 readers, excluding practice and repeated-reading trials, and treat original and simplified paragraph versions as distinct texts. This yields 126 texts and 9,718 readings. OneStop provides the same six per-word gaze measures used for ZuCo (Table 8); we copy each word's gaze vector to all corresponding GPT-2 sub-tokens, following the ZuCo pipeline. We retain GPT-2 Large as the frozen prior and leave all other components unchanged. The main analysis uses 1,000 permutation draws and the structure-only control uses 300.

Table 8: Mapping from OneStop interest-area columns to the six gaze quantities used for ZuCo.

GAZE QUANTITY	ONESTOP COLUMN	MEANING
fixated	IA_FIXATION_COUNT > 0	was the word looked at?
$\log(1 + n_{\text{Fix}})$	IA_FIXATION_COUNT	number of fixations
$\log(1 + \text{FFD})$	IA_FIRST_FIXATION_DURATION	first-fixation duration
$\log(1 + \text{GD})$	IA_FIRST_RUN_DWELL_TIME	gaze duration (first pass)
$\log(1 + \text{TRT})$	IA_DWELL_TIME	total reading time
$\log(1 + \text{GoPT})$	IA_REGRESSION_PATH_DURATION	go-past time

Information bound. On the validation measurement half (42,240 tokens from 14 texts), real gaze gives 0.1138 bits/token (DV bound 0.1071; $p = 0.001$ under both nulls). With a text-clustered bootstrap (10,000 resamples over texts), the bound is 0.114 [0.100, 0.127] under the derangement null and 0.116 [0.104, 0.128] under the roll null, close to ZuCo's 0.107 [0.090, 0.123] and 0.104 [0.091, 0.119]. The structure-only control (each word's gaze vector replaced by that of a random word, keeping sentence length, word grouping, and sub-token copying) gives 0.0911 bits/token (DV 0.0845; $p = 0.0033$ under both nulls), about 80% of the real value, compared with about 10% on ZuCo. Most of the OneStop bound therefore reflects text layout rather than gaze content; we did not determine why the layout signal is so much stronger on OneStop.

Word-level view. Because a word's gaze vector is copied to each of its sub-tokens, we also report the bound per word and for one-token words alone, where nothing is repeated (Table 9; a re-analysis of the stored evaluation draws with no new training). On one-token words OneStop gives 0.052 bits/token [0.042, 0.061], about half of ZuCo's 0.099 [0.077, 0.122], while multi-token words carry much more (0.226 vs. 0.119 on ZuCo). The OneStop bound is thus concentrated in multi-token words, consistent with the layout effect; the positive one-token value indicates that some gaze-content information remains.

Table 9: OneStop gaze at word level (measurement half, derangement null; roll null in the last column). Per-word bits are the sum over a word's sub-tokens. CIs are text-clustered bootstrap.

SUBSET	WORDS	TOKENS	BITS/TOKEN [95% CI]	BITS/WORD [95% CI]	BITS/TOKEN (ROLL)
All words	33,209	42,240	0.114 [0.100, 0.127]	0.145 [0.125, 0.164]	0.116
One-token words	27,209	27,209	0.052 [0.042, 0.061]	0.052 [0.042, 0.061]	0.049
Multi-token words	6,000	15,031	0.226 [0.188, 0.260]	0.566 [0.470, 0.654]	0.237
first sub-token	–	6,000	0.160 [0.115, 0.207]	–	0.164
continuation sub-tokens	–	9,031	0.269 [0.194, 0.330]	–	0.286

Selection. Table 10 gives validation selection accuracy. Real gaze is far above its controls, but the validation split has few texts, so pools stop at $N = 16$ and accuracies are coarse, and the text-only attack exceeds chance by more than two points at $N = 2$, 4, and 16. We report these numbers for information only and run no test session on OneStop.

Table 10: OneStop gaze: validation selection accuracy (same protocol as Appendix E). Attack = best text-only rule on the same pools.

N	CHANCE	REAL	ZEROED	$\gamma = 0$	DERANG.	ATTACK / GATE
2	0.500	0.971	0.500	0.412	0.441	0.647 / fail
4	0.250	1.000	0.265	0.206	0.147	0.294 / fail
8	0.125	1.000	0.176	0.147	0.029	0.118 / pass
16	0.062	0.912	0.088	0.088	0.029	0.118 / fail

Scope. OneStop has gaze only, far fewer unique texts than ZuCo (126 vs. 1,319), and a measurement half of only 14 texts. It supports the existence of a gaze signal beyond the prior on a second corpus with a much larger reader population, but not its exact size.

D Instrument Calibration Results (E1)

Table 11 presents the three-seed low-capacity follow-up calibration at the paper's fixed evaluation gain $\gamma_{\text{eval}} = 0.127$. The experiment is a sensitivity analysis of the already specified estimator: no recovered value exceeds the injected capacity, but reliable detection is not observed for weak synthetic channels.

E Selection Accuracy Results (E3)

E.1 Evaluation Protocol

The selection test asks whether evidence recorded while reading one sentence can identify that sentence among N candidates.

Evidence and pools. For each unique text, we construct one pool using its first recorded reading, so all texts receive equal weight. We pair the target with $N - 1$ distractors from other texts in the same split, never from training. We first restrict distractors to texts whose GPT-2 token length is within ± 2 of the target, retain the $\max(N-1, 4N)$ candidates with the closest mean per-token prior log-probability, and randomly sample $N - 1$. We also randomize the target position. If fewer than $N - 1$ texts satisfy the length constraint, we sample from all other texts rather than reduce N. We hash and log every pool.

Scoring and decision. We score each candidate y using the target evidence, $S(y) = \sum_t T_t(y_t, E_t)$ (equation 6), which measures how much the evidence raises its probability above the frozen prior. We pair candidate token t with evidence position t; because candidate lengths differ by at most two tokens, slightly longer candidates reuse the first evidence positions for their final tokens. We select the highest-scoring candidate and break exact ties uniformly at random.

Accuracy, chance, and bits. Top-1 accuracy P is the fraction of pools for which the target is selected, with chance $1/N$. We sweep $N \in \{2, 4, 8, 16, 32, 64\}$ over both validation halves, using one pool per unique text. We select checkpoints using the InfoNCE estimate on the selection half, never selection accuracy. We convert accuracy to Fano bits, $B = \log_2 N + P \log_2 P + (1 - P) \log_2 \frac{1-P}{N-1}$, and set $B = 0$ when $P \leq 1/N$.

Table 11: Low-bit synthetic calibration at the fixed evaluation gain $\gamma_{\text{eval}} = 0.127$. Detection requires $p < 0.05$ under both nulls.

INJECTED b	RECOVERED $\hat{I}$ (MEAN $\pm$ SD)	DETECTED SEEDS	MAX $\hat{I}$
0.000	-0.0022 ± 0.0010	0/3	-0.001
0.005	-0.0022 ± 0.0026	1/3	$+0.001$
0.010	-0.0047 ± 0.0045	0/3	-0.001
0.025	-0.0008 ± 0.0010	0/3	$+0.001$
0.050	-0.0031 ± 0.0030	0/3	$+0.000$
0.100	-0.0023 ± 0.0028	0/3	$+0.000$
0.150	-0.0007 ± 0.0005	0/3	-0.000
0.250	$+0.0026 \pm 0.0026$	2/3	$+0.006$
0.500	$+0.0294 \pm 0.0103$	3/3	$+0.043$
1.000	$+0.0771 \pm 0.0154$	3/3	$+0.093$
2.000	$+0.0951 \pm 0.0204$	3/3	$+0.121$

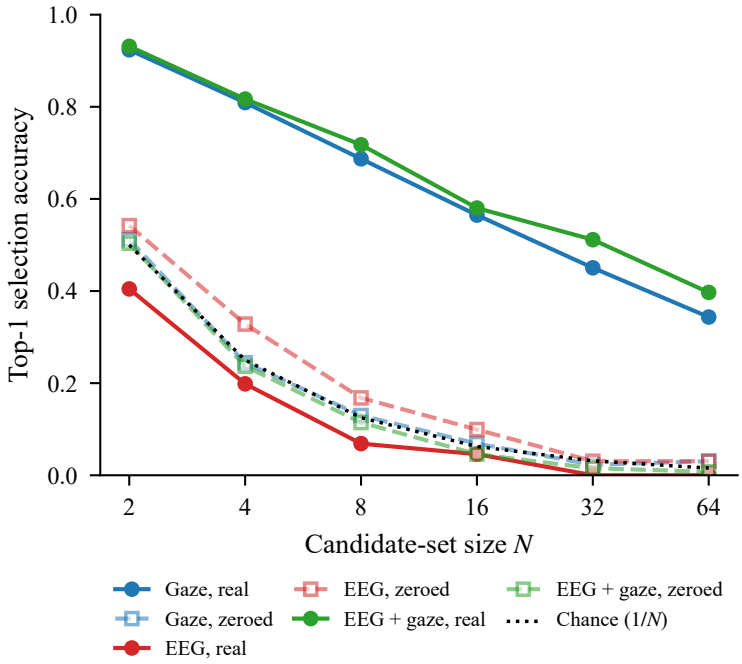


Figure 4: Selection accuracy vs. candidate pool size N (validation). Gaze scales with N; EEG shows no evidence-dependent advantage over its controls.

Control arms. We re-score the same model and pools under three evidence controls: *zeroed* evidence, $\gamma = 0$ (all candidates tie), and *derangement*, where each target receives evidence from another text in the split. Performance above these controls indicates dependence on the evidence rather than the prior or pool construction.

Pool-validity gate. To test for text-only shortcuts, we attack each pool set with six evidence-free rules: select the longest, shortest, median-length, highest type–token-ratio, highest-prior, or lowest-prior candidate. We use the best attacker accuracy and require it to remain within two percentage points of chance. If a pool set fails, we rebuild it once with a new seed while retaining the failed attempt in the log; a second failure marks the stage as failed. The gate passes on validation for $N = 32$ and 64 but fails for the small test pools used in E4 ($N = 2, 4$; Section 5.3).

Test evaluation (E4). We construct test pools identically at the frozen size N^*, yielding 133 pools, one per test text. We do not rebuild them after the validity check; instead, we log and disclose the outcome and any override. E4 also includes three additional control arms (Appendix G).

Figure 4 and Table 12 present validation selection accuracy across $N \in \{2, 4, 8, 16, 32, 64\}$. Gaze decreases predictably from 92.4% at $N = 2$ to 34.4% at $N = 64$. The scaling analysis in Section 5.6 retrains the gaze model separately and reaches 42.0% at $N = 64$ with the full training set; the two values come from different training runs of the same configuration. EEG provides no evidence-dependent advantage across these pool sizes and is at or below chance in the real-evidence arm.

Table 12: E3 top-1 selection accuracy on validation across candidate pool sizes. The subscript 0 denotes the zeroed-evidence arm.

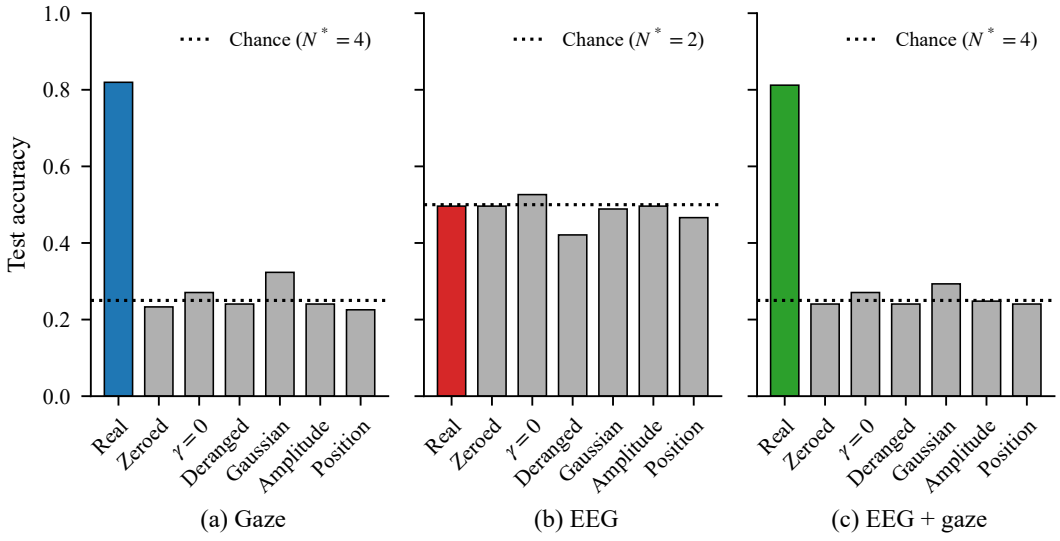
N	CHANCE	GAZE	GAZE_0	EEG	EEG_0	EEG+GAZE	$(\text{EEG+GAZE})_0$
2	0.500	**0.924**	0.511	0.405	0.542	**0.931**	0.504
4	0.250	**0.809**	0.244	0.198	0.328	**0.817**	0.237
8	0.125	**0.687**	0.130	0.069	0.168	**0.718**	0.115
16	0.062	**0.565**	0.069	0.046	0.099	**0.580**	0.046
32	0.031	**0.450**	0.023	0.000	0.031	**0.511**	0.015
64	0.016	**0.344**	0.031	0.000	0.031	**0.397**	0.008

F Complete Guarded Test Results (E4)

While the communication-system metrics for E4 are reported in the main text (Table 2), Table 13 and Figure 5 provide the complete breakdown of the single guarded test session (133 texts) across all seven arms (real evidence and six controls). Figure 6 shows the gaze ITR on the BCI axis and the test risk–coverage curve.

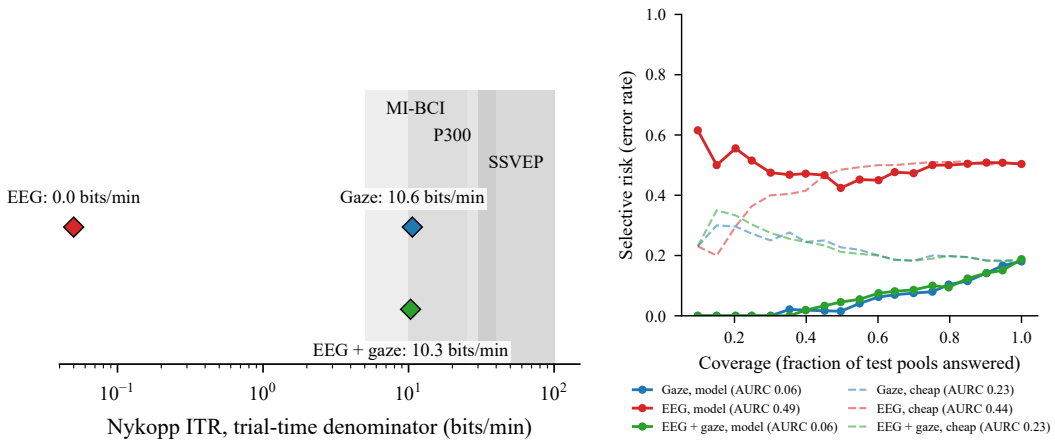
Table 13: E4 logged test session (133 pools) under the disclosed pool-validity override. Values are for seed 0 (gaze mean over three seeds: 0.832). Gaze is well above its control arms; EEG is evaluated at $N = 2$ and shows no observed evidence-dependent gain.

ARM	GAZE ($N = 4$)	EEG ($N = 2$)	EEG+GAZE ($N = 4$)
Chance	0.250	0.500	0.250
Real evidence	**0.820**	**0.496**	**0.812**
Zeroed evidence	0.233	0.496	0.241
$\gamma = 0$ (tie)	0.271	0.526	0.271
Text-deranged	0.241	0.421	0.241
Gaussian-matched	0.323	0.489	0.293
Amplitude-only	0.241	0.496	0.248
Position-only	0.226	0.466	0.241



(a) Gaze (b) EEG (c) EEG + gaze

Figure 5: E4 control battery under the logged pool-validity override. The gaze real-evidence arm is well separated from its no-evidence controls; EEG at $N = 2$ shows no observed advantage over the zeroed/control arms.



(a) ITR on the classical BCI axis

(b) Risk–coverage (test)

Figure 6: Communication-system metrics under the logged candidate-pool protocol. Gaze provides positive selection throughput and useful abstention; these values should not be interpreted as a demonstrated free-form brain–computer-interface (BCI) speller (Wolpaw et al., 2002; Chen et al., 2015). Values are for seed 0 (gaze three-seed mean: 11.1 ± 0.6 bits/min); because errors are spread uniformly over the wrong candidates, the Nykopp and Wolpaw ITRs coincide.

G Additional Control Analyses

G.1 Zero-Anchor Verification

For all runs, setting $\gamma = 0$ produces the "tie" condition in which the model asserts no evidence dependence. The resulting zero anchor has a spread of at most $6 \cdot 10^{-15}$ nats across all tokens, confirming the numerical stability of the exact partition-function implementation.

G.2 Test-Pool Attack Metrics

The "no-brain" attack, which selects candidates using sentence length, type–token ratio, or prior log-probability alone, achieves 30.8% at $N = 4$ compared with 25.0% chance and 56.4% at $N = 2$ compared with 50.0% chance. Both exceed the pre-specified chance $+2$ percentage-point criterion and therefore require the logged override used for the small-N E4 evaluation. In contrast, the attack remains within the specified margin at $N = 32$ and $N = 64$. These results quantify the text-only bias in the candidate pools and explain why the small-N E4 results are interpreted as controlled diagnostic evidence rather than as a clean pool-validity pass.

G.3 Control-Arm Breakdown

Table 13 reports the complete control battery. We evaluate six no-evidence controls. *Zeroed* sets all evidence values to zero; $\gamma = 0$ disables the evidence-dependent tilt, causing all candidates to tie; and *derangement* assigns each target evidence from a different randomly selected text, aligned by position. *Gaussian-matched* replaces evidence with noise matched to each feature's mean and standard deviation. *Amplitude-only* repeats the sentence-level mean evidence vector at every position, preserving magnitude but removing position-specific content, whereas *position-only* uses a fixed sinusoidal positional code without measured evidence.

With gaze, real evidence substantially outperforms all six controls, which cluster near chance ($N = 4$). For EEG, real and control performance cluster near chance ($N = 2$). Because the TOST analysis is underpowered, we interpret these comparisons as evidence-dependence diagnostics rather than formal evidence of equivalence to chance.

G.4 Structure-Only Control and Segmentation Patterns

The encoder processes a window of ± 1 *token*, and sub-token replication means that pieces of the same word inherit identical evidence rows. To test whether this segmentation structure alone explains the gaze result, we retrain after replacing each word's 6-d gaze vector by the gaze vector of a uniformly sampled word from the split while preserving sentence length, word grouping, and sub-token replication. The resulting structure-only lower bound is 0.0112 bits/token (DV 0.0050; $p = 0.0033$ under both derangement and roll), compared with 0.1065 bits/token for real gaze. Thus, segmentation/length structure produces a small but nonzero certified signal, while the real-gaze lower bound is substantially larger. Because both values are variational lower bounds, their difference is not treated as an exact decomposition of true mutual information.

G.5 EEG Representation and Positive-Control Analyses

To test whether the EEG result depends on the particular temporal alignment, we repeat E2 using the current-word window, a ± 3-word window, first-fixation band power, and total-reading-time band power. All variants yield near-zero information-bound estimates (Table 14). We additionally inject a controlled gaze-derived signal into the EEG features: α times a fixed random $6 \to 840$ projection of standardized gaze, scaled by each EEG feature's standard deviation. The weakest injection ($\alpha = 0.1$) remains undetected, whereas stronger injections ($\alpha = 0.3$ and 1.0) pass both null tests. These controls further restrict the EEG conclusion to effects detectable by the present representation and estimator.

Table 14: EEG representation and spike-in controls. GD: gaze-duration EEG window. Non-main runs use 300 permutation draws.

CONDITION	**INFONCE**	**DV**	p **(DER./ROLL)**
Main EEG, GD, ± 1	0.0000	-0.0000	0.270 / 0.254
Current word only	-0.0000	-0.0000	0.844 / 0.794
± 3-word window	-0.0000	-0.0250	1.000 / 1.000
First-fixation EEG	-0.0000	-0.0318	1.000 / 1.000
Total-reading-time EEG	-0.0011	-0.0988	1.000 / 1.000
Spike-in, $\alpha = 0.1$	-0.0395	-0.0837	0.399 / 0.083
Spike-in, $\alpha = 0.3$	0.0109	0.0071	0.003 / 0.003
Spike-in, $\alpha = 1.0$	0.0095	-0.0076	0.003 / 0.003

G.6 Leave-One-Subject-Out Generalization

Table 15 gives the per-reader results underlying the cross-reader analysis in Section 5.6. For each of the 12 ZuCo 1.0 readers, all readings from that reader are excluded from training and selection is evaluated on the held-out reader at $N = 4$ and $N = 32$. "Zeroed" denotes the corresponding no-evidence control.

Table 15: Leave-one-subject-out gaze selection on ZuCo 1.0. Chance accuracy is 0.250 at $N = 4$ and 0.031 at $N = 32$.

HELD-OUT READER	POOLS	ACC. N=4	ZEROED	ACC. N=32	ZEROED
ZAB	101	0.871	0.277	0.515	0.030
ZDM	101	0.723	0.267	0.436	0.030
ZDN	94	0.691	0.298	0.309	0.032
ZGW	95	0.789	0.211	0.474	0.021
ZJM	101	0.871	0.218	0.624	0.040
ZJN	101	0.881	0.267	0.525	0.020
ZJS	96	0.740	0.281	0.396	0.031
ZKB	96	0.698	0.219	0.354	0.021
ZKH	101	0.851	0.248	0.545	0.030
ZKW	101	0.832	0.238	0.525	0.010
ZMG	101	0.802	0.257	0.455	0.030
ZPH	88	0.750	0.295	0.432	0.011
Mean (12 readers)	–	**0.792**	0.256	**0.466**	0.025

H GENERATION BASELINE REPRODUCTION

To contextualize our findings within the existing EEG-to-text literature, we ported the Wang & Ji BrainTranslator model to our preprocessing pipeline and text-clustered split, evaluating it under free-running decoding (no teacher forcing). Table 16 presents results under both their original two-step SGD recipe and a modern AdamW recipe.

Under both optimization recipes, the generated outputs for real, shuffled, and noise EEG inputs are identical to three decimal places. Thus, in this reproduction, the generated text is effectively insensitive to the supplied EEG input. The BLEU-1 scores of 13.04 with SGD and 19.17 with AdamW are also comparable to the corrected literature baseline of 13.7 reported by Jo et al. (2024). In contrast, the gaze selection experiment at $N = 4$ yields a BLEU-4 score of 85.4 for the selected candidate with real evidence versus 24.8 with zeroed evidence, consistent with substantial evidence dependence in this controlled selection setting.

Table 16: Generation baselines using the Wang & Ji model under our preprocessing/split and free-running decoding. Identical values across arms are exact, not rounded.

MODEL / RECIPE	ARM	BLEU-1	BLEU-4	CHRF	ROUGE-1	BERTSCORE-F1	WER
BrainTranslator, their recipe	Real	13.04	0.60	8.8	0.106	0.851	0.99
	Shuffled / Noise	13.04	0.60	8.8	0.106	0.851	0.99
BrainTranslator, AdamW	Real	19.17	0.07	18.8	0.160	0.844	1.25
	Shuffled / Noise	19.17	0.07	18.8	0.160	0.844	1.25
Gaze selection ($N = 4$, ours)†	Real / Zeroed	86.9 / 33.1	85.4 / 24.8	87.7 / —	0.83 / —	0.97 / 0.88	0.16 / —

†Selected-text metrics depend on selection accuracy and are shown only to place the two paradigms on a common scale.

I TRAINING DETAILS

I.1 TRAINING CONFIGURATION

We train the evidence encoder using the InfoNCE objective (van den Oord et al., 2018), optimizing a contrastive lower-bound surrogate by increasing scores of matched word–evidence pairs relative to randomly permuted negatives. Optimization is performed using AdamW (Loshchilov & Hutter, 2019) with a learning rate of 10^{-3}, weight decay of 0.01, and a cosine learning rate schedule with a 5% warmup. Training runs for up to 20 epochs with early stopping patience of 5 epochs, a batch size of 8 sentences, and $M_{\text{tr}} = 4$ in-batch deranged negative evidence vectors per token. All hyper-

Table 17: Complete working hyperparameters. These settings are identical for the synthetic calibration and all three ZuCo evidence channels. Key: r, dimension of u_t; d_s, per-band projection dimension; d_{model}, Transformer width; $\sigma_{\max}$, spectral-norm bound on B; $\gamma_{\max}$, gain cap; SDPA, scaled dot-product attention; wpm, words per minute.

GROUP / PARAMETER	VALUE
Prior: model / vocabulary / embedding dim	gpt2-large / 50,257 / 1,280
Encoder: window / r / d_s / d_{model} / heads / layers / dropout	± 1 token / 16 / 64 / 128 / 4 / 2 / 0.1
Encoder: $s_{\max}$ (bound on $\|u_t\|$) / centring	4.0 / batch mean, running mean at eval
Tilt: $\sigma_{\max}(\|B\|_2)$ / $\gamma_{\max}$ / γ gate	1.0 / 5.0 / constant 1 in training
Tilt: nonlinear term φ	MLP $1280 \to 64 \to 16$, GELU, vocabulary-centred
Objective: loss / negatives per token	InfoNCE / 4 in-batch derangements
Optimiser: AdamW lr / weight decay / schedule / clip	10^{-3} / 0.01 / 5% warm-up + cosine / 1.0
Training: batch (sentences) / epochs / patience / selection stat.	8 / 20 / 5 / InfoNCE on selection half
Estimation: perms per null / bootstrap re-samples / nulls	1,000 / 10,000 / derangement, roll
Calibration follow-up: capacities / seeds / permutations	$\{0, .005, .01, .025, .05, .1, .15, .25, .5, 1, 2\}$ / 3 / 300
Selection: N grid / pool matching / attack margin	$\{2, 4, 8, 16, 32, 64\}$ / length ± 2, prior log-p / 2 pp
Test: pools / arms / ITR denominators / TOST margin	133 / 7 / T1, T2 (= T1 + 2 s), T3 (240 wpm) / ± 2 pp
Determinism: cuDNN / math / algorithms	deterministic, SDPA, `use_deterministic_algorithms`

parameters are finalized on synthetic calibration channels and are never tuned on the ZuCo test set (see Appendix Table 17 for complete architectural details).

I.2 Evaluation Metrics and Communication-System Analysis

We report four complementary operational quantities:

1. **Information Estimate:** Calibrated InfoNCE estimate $\hat{I}$ in bits/token.
2. **Selection Efficiency:** Fano bits (B) and Information Transfer Rate (ITR) in bits/min using the Wolpaw formulation (Wolpaw et al., 2002) across standardized trial durations T_1, T_2, and T_3.
3. **Abstention:** Area under the risk–coverage curve (AURC) using top-1/top-2 score margins, evaluated via confidence AUC and partial AUC (pAUC) after regressing out confounds.
4. **Equivalence Testing:** Two One-Sided Tests (TOST) (Lakens, 2017) to assess control-arm accuracy against chance.

Statistical inference. One-sided permutation p-values ($K = 1{,}000$) (Nichols & Holmes, 2002) and 95% confidence intervals from 10,000 bootstrap resamples are reported. Pre-registered hypotheses (**H1**–**H5**) govern significance testing without omnibus pooling.

I.3 Hyperparameter Configuration

Table 17 lists every hyperparameter used in our validated instrument. These settings are identical for the synthetic calibration channel and all three ZuCo evidence channels (EEG, Gaze, and EEG+Gaze), ensuring a strictly fair and controlled comparison across modalities; for COFETT, only the prior row changes (Qwen2.5-0.5B / 151,936 / 896, so φ is $896 \to 64 \to 16$). No hyperparameter was tuned on the test split; all choices were finalized on the synthetic calibration channel (Stage E1).

J TEST SPLIT ACCESS LOG AND REPRODUCIBILITY

The test split (fingerprint `3d3c89dbba05437d`, 133 unique texts, 1,925 readings) was accessed through a guarded, logged loader. Channel-specific pool sizes were frozen before test evaluation: $N^* = 4$ for gaze and EEG+gaze and $N^* = 2$ for EEG. The evaluation campaign comprised the original E1 calibration at $b \in \{0, 0.25, 1, 2\}$ bits/token, followed after the main campaign by the synthetic-only low-capacity sweep over $b \in \{0, 0.005, 0.01, 0.025, 0.05, 0.10, 0.15, 0.25, 0.5, 1, 2\}$; this follow-up did not access the biological test split. E2 channel measurements were performed on the validation measurement half (970 readings of 68 texts) for gaze, EEG, and EEG+gaze. The gaze test was independently replicated across three seeds, yielding accuracies of 82.0%, 82.7%, and 85.0% at $N = 4$. Each channel was additionally evaluated using seven arms—real, zeroed, $\gamma = 0$, text-deranged, Gaussian-matched, amplitude-only, and position-only—together with ITR, AURC, partial-AUC, and TOST analyses. The Wang & Ji BrainTranslator reproduction was performed on an earlier build containing 1,831 test sentences. Every test access is logged with its configuration hash, InfoNCE estimand, evidence channel, seed, gate outcomes, and timestamp. The model architecture and optimization settings were fixed before biological test evaluation; the later low-bit calibration is reported only as a sensitivity analysis and is not used to retune the biological-channel models. No hyperparameter was tuned on the test split. All numbers are produced by `run_v2.sh` and `run_v2_followup.sh`, together with `run_additional_analyses.sh` for the leave-one-subject-out, per-subgroup task/subject breakdown, instance-split leakage check, structure-only gaze control, attribute probe, generation-baseline reproduction, and low-bit calibration sweep, from the ZuCo pickles and `.mat` files; every reported value is a record in the ledgers of `runs_v2*/` with complete provenance. Figures are regenerated from the gate artifacts by `python -m cprd.pubfigs`.

K LIMITATIONS

Our findings should be interpreted within several limitations. First, the negative EEG result is specific to the evaluated fixation-locked, word-level band-power representation, estimator, and datasets; it should not be generalized to EEG-based language decoding in general. The synthetic calibration indicates limited sensitivity to weak channels, detecting $b = 0.25$ bits/token in only 2/3 seeds and $b \geq 0.5$ in 3/3 seeds. Second, the candidate-pool validity criteria were not satisfied at the selected pool sizes, requiring a disclosed pool-validity override; the 133 test pools also provide limited power for the TOST equivalence analysis. Third, LOSO generalization is established only across the 12 ZuCo 1.0 readers, while the COFETT analysis is limited to two participants, character-level features, and a different language prior. Finally, on OneStop, a substantial fraction of the measured gaze information is attributable to text layout, highlighting that a positive information bound does not necessarily imply uniquely cognitive or linguistic information. These limitations motivate broader validation across participants, datasets, neural representations, and stronger sensitivity regimes.