A COMMON MEASURE OF COMMUNICATION FOR SPEECH BRAIN–COMPUTER INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech decoding brain–computer interfaces (speech BCIs) translate neural activity into language, offering a path towards restoring speech for people with paralysis and, more broadly, enabling new forms of natural human–computer interaction. Despite this promise, the field lacks a common measure of progress because systems use different datasets, recording methods, types of speech, and vocabularies, so their reported scores are rarely comparable. Underlying this measurement problem are two unresolved questions: (i) what distribution of words should a speech BCI enable a user to communicate, and (ii) how much information from this distribution can a system convey. We address both by deriving *open-vocabulary mutual information* (OVMI), an information-theoretic quantity that measures the information conveyed by a decoder relative to a reference distribution over the words a user may wish to communicate. This allows capabilities measured under different conditions, such as distinct vocabularies, to be evaluated on a common reference scale. We show that ordinarily reported accuracy, word error rate (WER), and other metrics computed only over the words a system supports can overstate how much of a user's intended speech the system can communicate. We then use OVMI to compare existing systems, expose trade-offs between how much of the user's language a system supports and how accurately it decodes those words, show that these comparisons depend on what the user is expected to communicate, and demonstrate that selecting a vocabulary to maximise OVMI yields up to 16.3% relative improvement in accuracy across three speech domains. OVMI therefore provides the speech BCI community with a principled way to compare heterogeneous systems, improve vocabulary design, and measure progress in the field.

1 INTRODUCTION

In the latter half of the 1980s, DARPA proposed a speech recognition competition in which different systems transcribed the same held-out recordings and were ranked by a single score, settling the long-standing dispute between statistical methods and hand-written rules (Pallett, 2003; Donoho, 2017; Koch & Peterson, 2024). Standardised benchmarks built on this principle, with a shared task, dataset, and metric, have since underpinned progress across much of the machine learning field (Deng et al., 2009; Mnih et al., 2015; Wang et al., 2019). In the same vein, the field of speech BCIs has begun to adopt the same approach, with new benchmarks enabling controlled comparisons of methods that decode neural activity into speech (Willett et al., 2024; Card et al., 2025; Landau et al., 2025; Mantegna et al., 2026a). However, across the wider field, studies cannot all be evaluated with a common benchmark because methods are often specific to different experimental settings. Moreover, speech BCIs typically support different vocabularies of words, so studies report scores that are conditional on the supported vocabulary. As a result, scores from different studies are hard to compare.

If the vocabulary supported by a speech BCI is sufficiently large, then it can express most plausible sentences and the choice of supported words matters little. In practice, however, constrained vocabularies are more typical of current systems (Moses et al., 2021; Willett et al., 2023; Metzger et al., 2023; Özdogan et al., 2025; d'Ascoli et al., 2025; Jayalath & Parker Jones, 2026), where studies construct vocabularies in one of two ways. Some select the most frequent V words from the available data (d'Ascoli et al., 2025; Jayalath & Parker Jones, 2026); others curate a selection of words for a particular use case (Moses et al., 2021). Consequently, two systems with the same vocabulary size and accuracy may support very different fractions of what a user wishes to communicate. Conversely, imposing the

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.

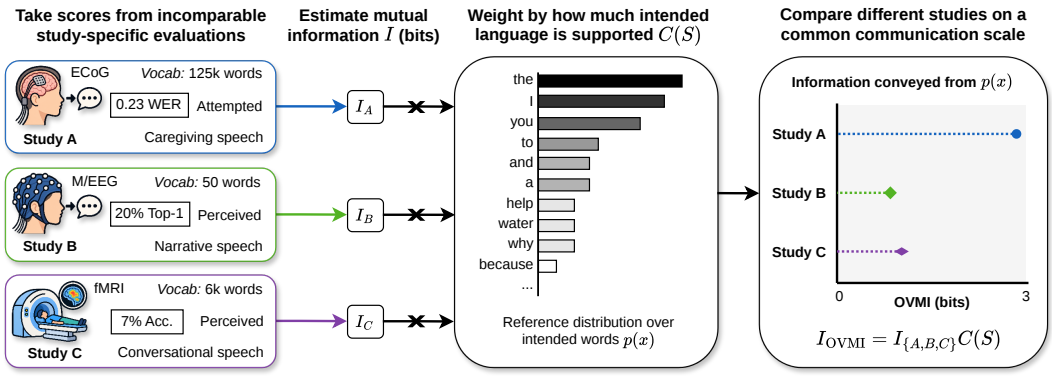

Figure 1: **OVMI maps study-specific evaluations onto a common reference scale.** Each study provides a score over its supported vocabulary S, from which we estimate the information conveyed I. We define a distribution $p(x)$ over the words a user may wish to communicate, and let $C(S)$ denote the probability that a word from $p(x)$ is in S. OVMI weights the decoded information I by $C(S)$, yielding a score relative to $p(x)$. Evaluating each measured lexical capability against the same $p(x)$ therefore places study-specific results on a common reference scale.

same vocabulary across datasets would create an unnatural and unfair comparison as those words may occur with very different frequencies, or not at all. Thus, neither standardising a universal vocabulary nor using existing study-specific vocabularies provide a satisfactory way to compare studies.

More fundamentally, no benchmark is able to capture the full range of speech BCI studies. Systems differ in recording method, from intracortical implants to non-invasive MEG and EEG; in speech task, e.g. attempted or perceived speech; in participant population, e.g. healthy volunteers or paralysed patients; and experimental protocol. These differences often make evaluation on a shared neural dataset impossible. Thus, benchmarks provide controlled comparisons of methods within an experimental setting but do not by themselves tell us how capabilities measured in different settings relate to one another. For example, an intracortical attempted-speech decoder and a non-invasive perceived-speech decoder cannot be compared as if they were evaluated under the same conditions, yet we may still wish to ask how much lexical decoding capability each demonstrates relative to the same communication objective. This question is also useful retrospectively, because many studies fall outside of the scope of a benchmark, and prospectively, since benchmarks may eventually saturate or be superseded. Therefore, a measure of communication should be meaningful across studies and successive benchmarks without depending on any one benchmark always staying relevant.

A benchmark also evaluates speech decoding on a particular language distribution. For example, the 2025 PNPL competition used Sherlock Holmes stories (Landau et al., 2025). Optimisation against such a benchmark may favour methods that decode only this language distribution well. We may instead want to evaluate the lexical capability measured on a benchmark relative to other communication distributions, such as conversational or caregiving speech. Collecting a new neural dataset for every such communication objective would be impractical. We therefore need to separate the language distribution used to measure decoding capability from the language distribution against which communication is assessed. Resolving these issues requires answering two underlying questions:

Unanswered questions for measuring progress in speech BCIs

1. What distribution (over words) should a speech BCI enable a user to communicate?
2. How much information from that distribution can a speech BCI convey?

(1) Defining the communication distribution. A speech BCI should ideally allow a user to convey what they would otherwise say aloud. To illustrate the problem, a decoder that is perfectly accurate on a fifty-word vocabulary succeeds only insofar as those fifty words capture what the user wants to say, and a vocabulary suitable for conversation may be poorly suited to caregiving. Moreover, the words a user needs are not uniformly distributed. We therefore model the distribution of words a user wants to communicate as a reference distribution p. This makes the target communication domain explicit

and provides a common distribution against which any decoder can be assessed. Thus, the amount of information a decoder conveys depends on both the decoder and the distribution of intended messages.

(2) Measuring communication against a reference distribution. Having specified what a user may wish to communicate through p, we ask how much of that communication a decoder can convey. While accuracy and word error rate (WER) measure decoding fidelity, they remain conditional on the vocabulary supported by the decoder. Additionally, they do not quantify how much information has been conveyed as, for example, the same classification accuracy resolves very different amounts of uncertainty when choosing among a vocabulary of ten words versus one thousand. Mutual information provides a natural and comparable quantity by measuring how much observing the decoder output reduces uncertainty about the intended message.

In BCIs, the standard information-theoretic measure is Wolpaw's information transfer rate and its refinements (Wolpaw et al., 1998; 2002; Speier et al., 2013). These measures quantify information within a predefined set of symbols. This is natural for classical closed-symbol interfaces, such as character spellers (Farwell & Donchin, 1988), but not when a user may intend words outside of a decoder's vocabulary. We therefore derive open-vocabulary mutual information (OVMI), which evaluates lexical information when the intended word is drawn from p. As Section 3 shows, OVMI emerges naturally from the decomposition of mutual information where OVMI is in-vocabulary mutual information weighted by the lexical coverage of p, the probability that an intended word is supported. Evaluating systems against the same p then allows for comparison with OVMI (Figure 1).

Contributions. We (i) show conventional measures can overstate communication and derive OVMI, an information-theoretic measure of lexical information transfer relative to an explicit reference communication distribution (Sections 3–4); (ii) use OVMI to compare heterogeneous speech BCI systems on a common scale, revealing how much systems depend on lexical coverage, decoding fidelity, and the intended communication domain (Sections 5.1–5.2); and (iii) show that OVMI can also guide vocabulary selection, improving speech BCI accuracy across three speech domains (Section 5.3).

2 Information-Theoretic Evaluation in BCIs

The standard information metric in BCIs is the information transfer rate (ITR) popularised by Wolpaw et al. (1998; 2002) and quoted in recent work (Moses et al., 2021; Perkins et al., 2025; Neuralink Corporation, 2026). Although ITR is in bits per minute, Wolpaw's formulation computes information per trial and multiplies by the trial rate. We work at the per-trial level—modelling continuous decoding as a series of independent word-level decoding attempts—since our primary interest is the information conveyed by each decoding attempt independently of communication speed.

With notation summarised in Appendix A, consider a decoder with a finite vocabulary S of size V. On each trial, the user intends a symbol $X \in S$, and the decoder outputs a symbol $Y \in S$. The quantity of interest is the mutual information $I(X;Y)$, which measures how much observing the output Y tells us about the intended symbol X. Wolpaw's derivation assumes all intended symbols are equally likely, $p(x) = 1/V$, and the decoder has symmetric errors, meaning it is correct with probability P and otherwise distributes errors uniformly over the remaining $V - 1$ symbols. Under these assumptions, the mutual information is (with derivation in Appendix B)

$$I_{\text{Wolpaw}}(X;Y) = \log_2 V + P \log_2 P + (1-P) \log_2 \left(\frac{1-P}{V-1} \right). \tag{1}$$

The uniform prior assumption is restrictive for natural language, where words are non-uniform and Zipfian (Zipf, 1949). Speier et al. (2013) addressed the problem in P300 character spellers (Farwell & Donchin, 1988) by using empirical character frequencies. More recently, Antonello et al. (2024) proposed a method for information-based evaluation of continuous semantic decoding (Tang et al., 2023). However, these formulations still measure information within a predefined set of symbols and do not account for whether that set can represent all the symbols a user may wish to communicate.

3 Open-Vocabulary Mutual Information

A speech BCI can convey an intended word only if the word is supported by the decoder and can be decoded reliably. Conventional accuracy, WER, and in-vocabulary mutual information measure only

the latter, whether a decoder can reliably decode words, conditional on the intended word belonging to the decoder vocabulary S. This is appropriate for a classification task, but not for measuring communication when a user may wish to say words outside of the supported vocabulary S.

OVMI therefore evaluates the decoder relative to an external reference distribution p over the words a user may wish to communicate. Its central idea is to weight information conveyed among supported words by the probability that an intended word drawn from p is supported. Thus, two decoders with identical in-vocabulary performance can differ in communicative capability according to OVMI if their vocabularies cover different amounts of p. The toy example below illustrates a more extreme case.

Toy example. Suppose a user may intend any of 1,000 equally likely words.

	System A	**System B**
Vocabulary size	50	1,000
In-vocabulary accuracy	**100%**	50%
Lexical coverage	5%	**100%**
OVMI	**0.28 bits**	**3.98 bits**

Despite perfect in-vocabulary decoding, System A can represent only 5% of what the user may wish to say. System B is less accurate but supports every intended word. Accuracy therefore ranks A above B, whereas OVMI ranks B above A by accounting for representability and decoding fidelity.

3.1 Formal Definition

Let p be a reference distribution over a lexicon Ω, and let $S \subset \Omega$ denote the decoder vocabulary, with $|S| = V$. For $X \sim p$, define the *lexical coverage* $C(S) = \Pr(X \in S) = \sum_{w \in S} p(w)$, the probability that an intended word is supported by the decoder. For explicitness, we assume that if $X \notin S$, then the user *abstains* from decoding, i.e. does not attempt to think or say the word, and the observable output is $Y = \varnothing$. If $X \in S$, the decoder produces some $Y \in S$. To denote whether the intended word is supported, i.e. whether $X \in S$, we use the indicator $Z = \mathbf{1}[X \in S] = \mathbf{1}[Y \neq \varnothing]$.

Proposition 1 (Decomposition of mutual information)**.** *Under this model,*

$$I(X;Y) = H_2(C(S)) + I(X;Y \mid Z) = H_2(C(S)) + C(S)\, I(X;Y \mid X \in S), \tag{2}$$

where $H_2(p) = -p \log_2 p - (1-p)\log_2(1-p)$ is the binary entropy. Proof in Appendix C.

Definition 1 (Open-vocabulary mutual information)**.** *We define*

$$I_{\mathrm{OVMI}}(S) := I(X;Y \mid Z) = C(S)\, I(X;Y \mid X \in S). \tag{3}$$

Intuition. How much lexical information is transferred per word the user would want to say?

OVMI is a component of mutual information, stating the information conveyed among supported words, weighted by how often a word drawn from the reference distribution is supported. The remaining term $H_2(C(S))$ reveals whether the intended word lies inside the decoder vocabulary. We exclude it because it contains no information about a word's identity and counting it would assign information to detecting if a word is in the vocabulary even if the decoder is not able to decode the word.

Relationship to existing metrics. When $C(S) = 1$, OVMI reduces to ordinary mutual information within the decoder vocabulary. If the intended words are additionally uniform and decoding errors are symmetric, it reduces to the per-trial information quantity underlying Wolpaw's ITR (Wolpaw et al., 1998; 2002). For a trial rate r, rI_{OVMI} therefore gives the corresponding open-vocabulary information-transfer rate, with conventional Wolpaw ITR as a special case. Accuracy and WER likewise measure decoding performance within the evaluated vocabulary but do not account for the probability that an intended word is supported. We analyse the impact of this in Section 4.

3.2 ESTIMATING OVMI

When a confusion matrix of decoding errors is available, OVMI may be computed from the decoder's in-vocabulary channel, $K_S(y \mid x) = \Pr(Y = y \mid X = x)$, which we describe fully in Appendix C.3. We also provide an estimator when only accuracies are known for every supported word in Appendix C.6. When evaluating the measured channel under a different reference p, we hold K_S fixed, which is an assumption that we formalise in Appendix D.2 (channel invariance).

Scalar approximation for retrospective comparisons. Most speech BCI studies report only a scalar accuracy or WER, and large confusion matrices are often poorly estimated from limited test data. We therefore also provide a Wolpaw-like approximation in which every word is decoded correctly with probability P and errors are distributed uniformly. We define this next in Corollary 1, providing a mapping from the scalar performance statistics typically available in published studies.

Corollary 1 (Scalar OVMI estimator)**.** *Suppose that, for every intended word $x \in S$, the decoder outputs the correct word with probability P. Conditional on an error, it distributes the remaining probability uniformly across the other $V - 1$ words. Thus*

$$K_S(y \mid x) = \begin{cases} P, & y = x, \\ \dfrac{1-P}{V-1}, & y \neq x. \end{cases}$$

Under this model, the output distribution on S is

$$q_S(j) = P\, p_S(j) + \frac{1-P}{V-1}\big(1 - p_S(j)\big), \qquad j \in S, \tag{4}$$

where $p_S(j)$ is the normalised distribution p restricted to S. Consequently (with proof in Appendix C),

$$\widehat{I}^{\text{scalar}}_{\text{OVMI}}(S) = C(S)\Big[H(q_S) - h_V(P)\Big], \qquad \textit{where } h_V(P) \textit{ is the conditional entropy.} \tag{5}$$

Defining P. We instantiate the scalar P in Eq. 5 as macro accuracy P_{macro}, the average of per-word correctness probabilities over S. Using micro accuracy would undesirably weight classes according to the evaluation set's frequency distribution, whereas source weighting in OVMI should be determined by the external reference p. The non-uniform source distribution is already modelled by p and enters through $C(S)$, p_S, and the output distribution q_S. When only WER is reported, we use $P = \max(0, 1 - \text{WER})$ as a practical proxy for the scalar estimator. Because WER counts insertions, this is a conservative estimate of words recovered in the ground truth.

Choosing an estimator. For retrospective comparisons of prior work, we adopt the scalar approximation (with pseudocode in Appendix D.4) because prior work reports only a scalar performance metric and a vocabulary. When per-word accuracy estimates are available, OVMI should be computed using the word-specific variant in Appendix C.6; when a confusion matrix is available, it should be computed directly from Proposition 2 (statement and proof in Appendix C.3). In the non-invasive settings considered here, however, the $O(|S|^2)$ parameters of a confusion matrix are often poorly estimated under the long-tailed data of naturalistic neural recordings (Hamilton & Huth, 2018). We quantify the robustness of the scalar and per-word approximations in Appendix D.3.

4 COMMON METRICS OVERESTIMATE OPEN-VOCABULARY PERFORMANCE

Before comparing empirical speech decoders, we first isolate error introduced by the metric. Consider an ideal, noiseless decoder ($P = 1$) whose vocabulary S consists of the top-V most frequent words under the reference distribution. With Wolpaw's formulation, such a decoder conveys $\log_2 V$ bits because the V supported words are assumed equally likely. Accounting for their non-uniform frequencies reduces this to the in-vocabulary entropy $H(p_S)$, but this assumes that the words a user intends to communicate always lie within S. OVMI corrects this misalignment by accounting for the probability that an intended word is supported at all, giving $C(S)H(p_S)$ with an ideal, noiseless decoder.

Figure 2a shows that these quantities diverge substantially. Of the three, OVMI is the most conservative measure of information transfer, as $C(S)H(p_S) \leq H(p_S) \leq \log_2(V)$, with equality on the left

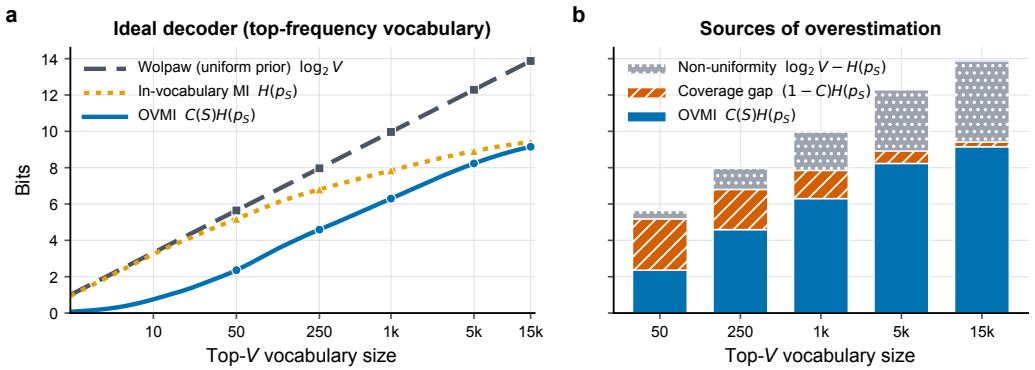

Figure 2: **In-vocabulary measures overestimate open communication. (a)** For a noiseless decoder supporting the most frequent words in p, Wolpaw assigns $\log_2 V$, non-uniform in-vocabulary MI gives $H(p_S)$, and OVMI gives $C(S)H(p_S)$. **(b)** The total error decomposes into missing coverage and error due to non-uniformity. The distribution p is computed from a spoken-language frequency norm derived from film and television subtitles (van Heuven et al., 2014).

when S provides perfect coverage, $C(S) = 1$, of the reference distribution, and equality on the right when S is uniformly distributed. Especially for small vocabularies, open-vocabulary metrics reveal that a system may convey only a small fraction of the information suggested by in-vocabulary entropy.

What causes overestimation? Figure 2b separates the discrepancy into its two sources. For a noiseless decoder, the Wolpaw uniform-prior quantity may be decomposed as

$$\log_2 V = \underbrace{C(S)H(p_S)}_{\text{OVMI}} + \underbrace{(1 - C(S))H(p_S)}_{\text{Coverage gap}} + \underbrace{[\log_2 V - H(p_S)]}_{\text{Non-uniformity}}. \tag{6}$$

The first term is OVMI, the second is excess due to evaluating in-vocabulary information without weighting by lexical coverage, and the third is error introduced by treating the supported words as uniformly distributed. At small V, in-vocabulary entropy and the uniform prior vastly overestimate information transfer, driven by incomplete lexical coverage, $C(S) \ll 1$, as even a perfectly accurate decoder cannot communicate an intended word that it does not support. As V increases, coverage approaches unity and OVMI approaches in-vocabulary entropy, however, the gap to the uniform-prior, $\log_2(V)$, continues to grow due to the Zipfian distribution of words in the reference p. Invasive systems (Moses et al., 2021; Willett et al., 2023) and recent non-invasive decoders (d'Ascoli et al., 2025; Jayalath & Parker Jones, 2026) use vocabularies of 50–250 Zipfian-distributed words, where both sources of overestimation are substantial.

Accuracy and WER overstate open-vocabulary capability. For a given vocabulary, Wolpaw's quantity is a monotonic transformation of in-vocabulary accuracy. Hence, accuracy inherits the issue of conditioning on this vocabulary. Under our model, the probability of decoding an intended word is

$$\Pr(Y = X) = \Pr(X \in S) \Pr(Y = X \mid X \in S) = C(S)P_{\text{in-vocab}}.$$

As a result, even perfect in-vocabulary accuracy can correspond to poor open-vocabulary communication when lexical coverage is low. Likewise, WER measures decoding errors on an evaluation transcript but does not by itself account for how much of an external reference distribution that system can represent. Consequently, high accuracy or low WER can occur at the same time as low OVMI when the vocabulary does not cover much of what a user may wish to communicate.

5 Results

We now use OVMI to evaluate existing speech BCI systems relative to an explicit communication distribution. Unless otherwise stated, we use SUBTLEX-UK (van Heuven et al., 2014) as p, treating its word frequencies from film and television subtitles as a broad spoken English reference. We first compare heterogeneous speech decoders with scalar OVMI (Section 5.1), then examine how the comparison changes with p (Section 5.2), and finally test full channel OVMI as an objective for vocabulary selection (Section 5.3). We provide practical details on estimating OVMI in Appendix E.

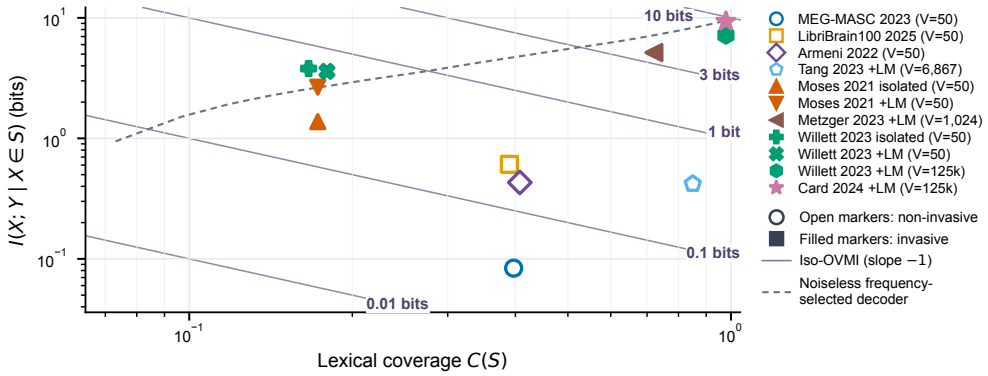

Figure 3: **Decomposition of OVMI into lexical coverage and decoding fidelity.** The product of the axes gives OVMI, with contours indicating equal OVMI. The dashed curve shows a noiseless decoder whose vocabulary consists of the V most frequent words under p. For frequency-selected vocabularies, vertical displacement from the noiseless curve reflects decoding error while movement leftwards reflects information foregone due to the limited coverage of smaller vocabularies.

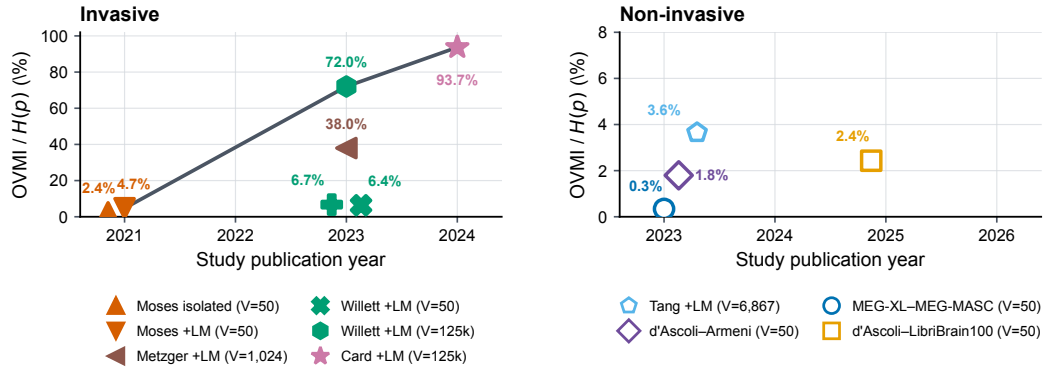

Figure 4: **Estimated speech BCI progress on a common reference scale.** Systems plotted by publication year using OVMI normalised by the entropy of p. The solid line connects the highest-scoring invasive system in each year; non-invasive results are shown unconnected because they correspond to datasets and not methods. Full values and uncertainties are provided in Appendix E.6.

5.1 A Common Scale for Heterogeneous Speech BCIs

We compare speech decoding systems that differ in vocabulary, dataset, task, and recording modality on the same reference scale by evaluating each against the same distribution p. For invasive work, the dataset and the decoder are typically reported together. We consider several landmark invasive studies (Moses et al., 2021; Willett et al., 2023; Metzger et al., 2023; Card et al., 2024) which decode attempted speech from partially paralysed patients. For non-invasive work, datasets are often released independently and subsequently evaluated with multiple decoding methods, so we use the strongest evaluated decoder for each dataset. For MEG-MASC (Gwilliams et al., 2023), we train MEG-XL (Jayalath & Parker Jones, 2026), while for the other MEG datasets—LibriBrain100 (Mantegna et al., 2026b) and Armeni (Armeni et al., 2022)—we train the decoder of d'Ascoli et al. (2025). Tang et al. (2023) is included as a study in which the dataset and decoding method are reported together. In these datasets, healthy participants listen to or view naturalistic stimuli such as movies or spoken narratives.

Where is information lost? Figure 3 separates the two factors determining OVMI. The 50-word invasive systems (Moses et al., 2021; Willett et al., 2023) convey comparatively high information within their supported vocabularies but cover only a small fraction of broad spoken English. Most evaluated non-invasive systems achieve greater coverage through frequency-selected vocabularies, but substantially lower in-vocabulary information. The later 125k-word invasive systems (Willett et al., 2023; Card et al., 2024) occupy the upper-right of the plot, combining near-complete coverage with high decoding fidelity. For frequency-selected vocabularies, displacement below the noiseless

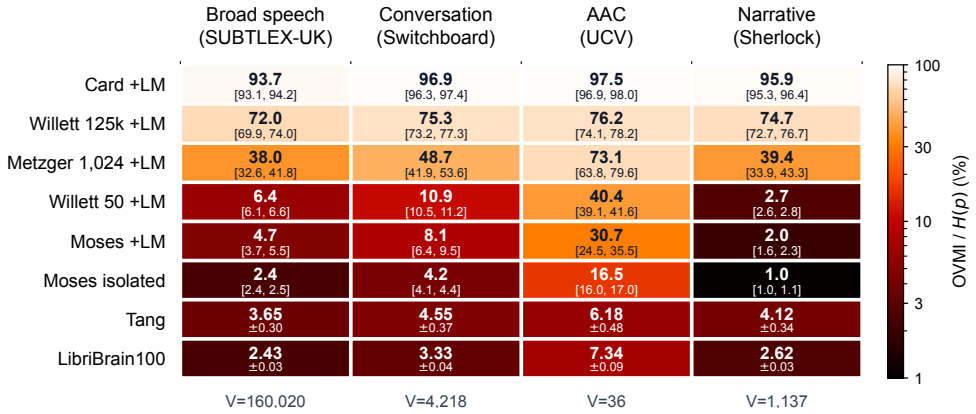

Figure 5: **OVMI depends on the reference distribution.** Cells report $I_{\mathrm{OVMI}}/H(p)$ under broad speech, conversation, AAC, and narrative reference distributions, with uncertainty beneath. Brackets denote propagated intervals for published invasive results, whereas $\pm$ denotes one-SEM uncertainty for Tang and LibriBrain100. Others excluded for brevity.

frontier reflects information lost through decoding error, while movement leftwards reflects limited lexical coverage. Curated vocabularies do not need to lie below this frontier because they may trade coverage for greater in-vocabulary entropy. The large-vocabulary invasive systems improve primarily by expanding coverage after decoding fidelity had already become strong, whereas the evaluated non-invasive systems remain limited by decoding fidelity, likely as a consequence of the lower signal fidelity afforded by sensors outside the skull.

Understanding the progress of systems over time. Figure 4 shows that in 2021, the 50-word Moses system conveys 4.7% or 2.4% of the lexical information available with and without language-model post-processing. The 2023 50-word Willett results remain in the same regime, conveying 6.4% and 6.7% respectively. The major change comes from increasing the vocabulary with Metzger's 1k-word system at 38.0%, then Willett's 125k-word system reaching 72.0%, followed by Card at 93.7% in 2024. Thus, the largest historical increase among invasive systems occurs with the transition to large-vocabulary decoding, after in-vocabulary performance had saturated with small vocabularies.

Non-invasive word decoding is in a lower information regime. The d'Ascoli et al. (2025) decoder yields 2.4% on LibriBrain100 and 1.8% on Armeni, MEG-XL yields 0.3% on MEG-MASC, and Tang's fMRI system reaches 3.6%. The low MEG-MASC value occurs in a challenging regime with little data per participant across a large subject population. For perceived speech, OVMI should be interpreted as lexical decoding capability relative to the speech distribution, rather than communication. When comparing across settings, it is important to note the underlying paradigm; for example, perceived speech decoding does not by itself imply a useful communication interface.

5.2 OVMI Depends on the Reference Distribution

OVMI is meaningful only relative to a specified communication target. Figure 5 evaluates the same systems with four reference distributions representing broad spoken English (SUBTLEX-UK; van Heuven et al., 2014), conversational speech (Switchboard; Godfrey et al., 1992), augmentative and alternative communication (AAC/UCV; Erickson et al., 2019), and narrative speech (Sherlock Holmes; Doyle, 1892). Because these distributions have different entropies, we normalise OVMI by $H(p)$.

If information transfer in OVMI were a property of the decoder alone, then there would be no difference between the columns. Instead, we find that large-vocabulary systems are comparatively insensitive to the choice of p as Card conveys 93.7–97.5% across the four references and the 125k-word Willett system 72.0–76.2%. However, systems with smaller vocabularies are much more sensitive. The 50-word Willett system, for example, conveys 6.4% of broad spoken-English information but 40.4% under the AAC reference and Moses similarly increases from 4.7% to 30.7%. These systems use vocabularies designed around clinical communication, and consequently cover a much larger proportion of the more restricted AAC distribution than of unrestricted speech. LibriBrain100 and the isolated-word Moses system are nearly matched under broad speech (2.43% versus 2.4%), but Moses

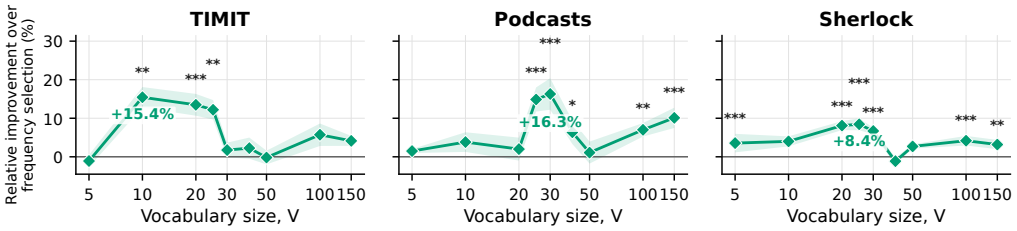

Figure 6: **OVMI-guided vocabulary selection improves open-vocabulary accuracy.** Shading denotes SEM over five seeds and stars denote paired permutation-test significance ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$). Absolute accuracies and further comparisons are provided in Appendix F.

leads substantially under AAC (16.5% versus 7.34%), whereas LibriBrain100 leads under narrative speech (2.62% versus 1.0%). Whether one system represents progress over another can therefore depend on what the interface is intended to communicate. Accordingly, OVMI should always be reported together with p and when uncertain, several plausible reference distributions should be reported.

5.3 OVMI as a Vocabulary Selection Objective

So far, we have used OVMI retrospectively on prior studies. We next ask whether it can also help to choose the vocabulary of a system. Recent contrastive decoders restrict a larger candidate lexicon to a smaller vocabulary at inference to improve accuracy (d'Ascoli et al., 2025; Jayalath & Parker Jones, 2026), creating a trade-off between supporting frequent or important words and retaining words that are easier to decode. OVMI provides a natural objective because it accounts for both of these factors.

We test this using the d'Ascoli et al. (2025) contrastive decoder. The model is trained once with a candidate lexicon of 250 words. At inference, decoding requires retrieving the most likely word from a set of candidate embeddings. For each vocabulary size V, we select a subset S_V and restrict retrieval to that subset, where varying V does not require any changes to the trained decoder. Figure 6 shows that selecting vocabularies to greedily maximise OVMI, using the training text as reference distribution and validation set decoder statistics, improves open-vocabulary accuracy (where out-of-vocabulary target words are counted as errors) over selecting the most frequent words, which is the strategy used in recent work (d'Ascoli et al., 2025; Jayalath et al., 2025; Jayalath & Parker Jones, 2026). We repeat this for three distributions: TIMIT (sentences covering different phonetic contexts), podcast conversations, and a Sherlock Holmes book chapter. We find that OVMI matches or exceeds frequency selection across vocabulary sizes, with peak relative improvements of 15.4%, 16.3%, and 8.4%. The advantage diminishes at larger vocabulary sizes as words are selected from a pool of 250, leading to convergence.

6 Discussion

Scope and limitations. (i) Our retrospective comparisons use the scalar approximation because confusion matrices are rarely reported, though we show this approximation generally preserves ranking (Appendix D.3). When richer statistics are known, OVMI can be computed from a confusion matrix (Appendix C.3) or per-word accuracies (Appendix C.6). (ii) OVMI penalises out-of-vocabulary words as unsupported, however, the same meaning may be paraphrased. We choose to keep the measure tied to the lexical information available through the vocabulary. (iii) We evaluate lexical not contextual information; extending OVMI to a conditional distribution is left to future work.

OVMI allows heterogeneous speech BCI systems to be compared by evaluating capabilities measured in different experimental settings against the same reference distribution. It complements benchmarks as these compare methods within a fixed setting, whereas OVMI relates capabilities across settings to the same communication objective. OVMI can also guide vocabulary design to improve accuracy. The comparisons in this work do not erase differences in experimental setting, and a higher OVMI does not necessarily imply a more useful paradigm; for example, perceived speech decoding on its own is unlikely to constitute a usable communication interface. We recommend reporting OVMI alongside other decoder metrics for future ease of comparison, and encourage user-centred evaluation before any practical use. We hope that OVMI's lasting role is to give the emerging field of speech BCIs a principled way to compare and improve different systems.

AI USE STATEMENT

Generative AI tools were used to help develop and refine some of the conceptual framework of this work, formulate and check mathematical claims, assist with derivations and proofs, refine research hypotheses, provide feedback on experimental methodology and design, and assist in interpreting experimental results. All AI-assisted theoretical claims, derivations, methodological decisions, and interpretations were independently reviewed and verified by the authors. The authors take responsibility for the final content of this work, including all claims and results produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve new human subject recruitment or data collection. Our experiments use previously collected neural datasets and published results from studies involving both healthy participants and people with paralysis, under the ethics and consent procedures reported by the original studies. Neural speech decoding raises important privacy and consent considerations because neural recordings may contain sensitive information. Our contribution is an evaluation framework rather than a new decoding method, and does not increase access to neural information. Nevertheless, we caution that the choice of reference distribution p in OVMI determines which intended language is valued by the metric and may inadequately represent particular users, languages, dialects, or communication needs. We therefore recommend reporting p explicitly and choosing it to reflect the intended users and application.

REPRODUCIBILITY STATEMENT

We provide the mathematical assumptions and definition of OVMI in Section 3, with complete derivations and proofs in Appendix C. The appendix also provides pseudocode for computing OVMI (Appendix D.4), alternative estimators using per-word accuracies and full confusion matrices, and full details of how reported accuracy and WER values from prior studies are converted to OVMI (Appendix E). Experimental details for the vocabulary selection experiments, including the evaluation protocol and statistical testing, are provided in Appendix F, and the complete cross-study results and uncertainties are reported in Appendix E.6. Code for computing OVMI, scripts for reproducing the analyses, and an interactive OVMI explorer are publicly available and will be linked in the camera-ready version following the anonymous review process.

REFERENCES

Richard Antonello, Nihita Sarma, Jerry Tang, Jiaru Song, and Alexander Huth. How many bytes can you take out of brain-to-text decoding? arXiv preprint arXiv:2405.14055, 2024.

Kristijan Armeni, Umut Güçlü, Marcel van Gerven, and Jan-Mathijs Schoffelen. A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. Scientific Data, 9(1):278, 2022. License: CC-BY-4.0.

Nicholas S. Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R. Willett, Erin M. Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R. Deo, Aparna Srinivasan, Eun Young Choi, Matthew F. Glasser, Leigh R. Hochberg, Jaimie M. Henderson, Kiarash Shahlaie, Sergey D. Stavisky, and David M. Brandman. An accurate and rapidly calibrating speech neuroprosthesis. New England Journal of Medicine, 391(7):609–618, 2024.

Nicholas S. Card, Maitreyee Wairagkar, Carrina Iacobacci, Xianda Hou, Tyler Singer-Clark, Francis R. Willett, Erin M. Kunz, Chaofei Fan, Maryam Vahdati Nia, Darrel R. Deo, Aparna Srinivasan, Eun Young Choi, Matthew F. Glasser, Leigh R. Hochberg, Jaimie M. Henderson, Kiarash Shahlaie, Sergey D. Stavisky, and David M. Brandman. Brain-to-text '25. `https://kaggle.com/competitions/brain-to-text-25`, 2025. Kaggle.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp. 248–255. IEEE Computer Society, 2009.

David L. Donoho. 50 years of Data Science. Journal of Computational and Graphical Statistics, 26: 745 – 766, 2017.

Arthur Conan Doyle. The Adventures of Sherlock Holmes. The Strand Magazine, 1892.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert J. Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Towards decoding individual words from non-invasive brain recordings. Nature Communications, 16(10521), 2025.

Karen Erickson, Lori Geist, Penny Hatch, and Nancy Quick. The universal core vocabulary. Technical report, Center for Literacy & Disability Studies, University of North Carolina at Chapel Hill, Chapel Hill, NC, 2019.

Lawrence A. Farwell and Emanuel Donchin. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography & Clinical Neurophysiology, 70(6):510–23, 1988.

John J. Godfrey, Edward Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1:517–520 vol.1, 1992.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing. Scientific Data, 10(1):862, 2023. License: CC0 1.0 Universal.

Liberty S. Hamilton and Alexander G. Huth. The revolution will not be controlled: Natural stimuli in speech neuroscience. Language, Cognition and Neuroscience, 35:573 – 582, 2018.

Dulhan Jayalath and Oiwi Parker Jones. MEG-XL: Data-efficient brain-to-text via long-context pre-training. International Conference on Machine Learning (ICML), 2026. arXiv preprint arXiv:2602.02494.

Dulhan Jayalath, Gilad Landau, and Oiwi Parker Jones. Unlocking non-invasive brain-to-text. International Conference on Machine Learning (ICML), Workshop on Generative AI and Biology, 2025. arXiv preprint arXiv:2505.13446.

Bernard J Koch and David Peterson. From protoscience to epistemic monoculture: How benchmarking set the stage for the deep learning revolution. arXiv preprint arXiv:2404.06647, 2024.

Gilad Landau, Miran Özdogan, Gereon Elvers, Francesco Mantegna, Pratik Somaiya, Dulhan Jayalath, Luisa Kurth, Teyun Kwon, Brendan Shillingford, Greg Farquhar, Minqi Jiang, Karim Jerbi, Hamza Abdelhedi, Yorguin Mantilla Ramos, Caglar Gulcehre, Mark Woolrich, Natalie Voets, and Oiwi Parker Jones. The 2025 PNPL competition: Speech detection and phoneme classification in the LibriBrain dataset. Advances in Neural Information Processing Systems (NeurIPS), Competition Track, 2025. arXiv preprint arXiv:2506.10165.

Francesco Mantegna, Gereon Elvers, Dulhan Jayalath, Gilad Landau, Tasha Kim, Miran Özdogan, Luisa Kurth, Teyun Kwon, SungJun Cho, Benjamin Ballyk, Alex Fung, Anna Greer, Pratik Somaiya, Christian Herff, Yorguin Mantilla Ramos, Hamza Abdelhedi, Karim Jerbi, Greg Farquhar, Brendan Shillingford, Mark Woolrich, and Oiwi Parker Jones. The 2026 PNPL competition: Word classification and efficient cross-subject generalisation in LibriBrain100. arXiv preprint arXiv:2609.03231, 2026a.

Francesco Mantegna, Dulhan Jayalath, Gereon Elvers, Tasha Kim, Benjamin Ballyk, Alex Fung, SungJun Cho, Teyun Kwon, Luisa Kurth, Miran Özdogan, Gilad Landau, Pratik Somaiya, Natalie Voets, Mark Woolrich, and Oiwi Parker Jones. LibriBrain100: One hundred hours of broad and deep MEG data for neural speech decoding at scale. arXiv preprint arXiv:2608.25204, 2026b.

Sean L. Metzger, Kaylo T Littlejohn, Alexander B. Silva, David A. Moses, Margaret P. Seaton, Ran Wang, Maximilian E. Dougherty, Jessie R. Liu, Peter Wu, Michael Berger, Inga Zhuravleva, Adelyn P. Tu-Chan, Karunesh Ganguly, Gopala Krishna Anumanchipalli, and Edward F. Chang. A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 620:1037 – 1046, 2023.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Kirkeby Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. Nature, 518:529–533, 2015.

David A. Moses, Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, Joseph G. Makin, Pengfei F. Sun, Josh Chartier, Maximilian E. Dougherty, Patricia M. Liu, Gary M. Abrams, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. New England Journal of Medicine, 385(3):217–227, 2021.

Neuralink Corporation. Two years of telepathy (article). `https://neuralink.com/updates/two-years-of-telepathy/`, 2026. Accessed: 2026-04-20.

David S. Pallett. A look at NIST's benchmark ASR tests: past, present, and future. 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721), pp. 483–488, 2003.

Sean M Perkins, Michael Trumpis, Michael E Reitman, Beata Jarosiewicz, Aashish N Patel, Adam Weiss, Jacob W Scott, Kurtis Nishimura, Matthew R Angle, Shaoyu Qiao, et al. SONIC: A benchmarking paradigm for brain-computer interfaces. bioRxiv preprint 2025.09.30.679683, 2025.

William Speier, Corey Arnold, and Nader Pouratian. Evaluating true BCI communication rate through mutual information and language models. PLoS ONE, 8(10):1–9, 10 2013.

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.

Walter J. B. van Heuven, Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. SUBTLEX-UK: A new and improved word frequency database for British English. Quarterly Journal of Experimental Psychology, 67:1176 – 1190, 2014.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019.

Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Guy H. Wilson, Eun Young Choi, Foram B. Kamdar, Matthew F. Glasser, Leigh R. Hochberg, Shaul Druckmann, Krishna V. Shenoy, and Jaimie M. Henderson. A high-performance speech neuroprosthesis. Nature, 620:1031 – 1036, 2023.

Francis R Willett, Jingyuan Li, Trung Le, Chaofei Fan, Mingfei Chen, Eli Shlizerman, Yue Chen, Xin Zheng, Tatsuo S Okubo, Tyler Benster, et al. Brain-to-Text benchmark '24: Lessons learned. arXiv preprint arXiv:2412.17227, 2024.

Jonathan R. Wolpaw, Herbert Ramoser, Dennis J. McFarland, and Gert Pfurtscheller. EEG-based communication: improved accuracy by response verification. IEEE Transactions on Rehabilitation Engineering, 6(3):326–33, 1998.

Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. Brain-computer interfaces for communication and control. Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology, 113(6):767–91, 2002.

George Kingsley Zipf. Human behavior and the principle of least effort. Addison-Wesley, 1949.

Miran Özdogan, Gilad Landau, Gereon Elvers, Dulhan Jayalath, Pratik Somaiya, Francesco Mantegna, Mark Woolrich, and Oiwi Parker Jones. LibriBrain: Over 50 hours of within-subject MEG to improve speech decoding methods at scale. Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2025. arXiv preprint arXiv:2506.02098.

APPENDIX CONTENTS

A Table of Variables **14**

B Derivation of Wolpaw's Information Transfer Rate **14**

C Proofs and Supplementary Derivations for OVMI **15**

C.1 Proof of Proposition 1 15

C.2 Equivalent Entropy Derivation of Proposition 1 16

C.3 Proposition 2 16

C.4 Proof of Proposition 2 16

C.5 Proof of Corollary 1 17

C.6 Retaining Per-Word Accuracies in OVMI 18

C.7 Further Observations for the Homogeneous Symmetric Channel 19

D Further Notes on OVMI **19**

D.1 Relation to Continuous Sequence Decoding 19

D.2 Reference Reweighting and Channel Invariance 19

D.3 Robustness of OVMI Estimators 20

D.4 Closed Form OVMI Pseudocode 21

E Cross-Study OVMI Estimation Details **21**

E.1 Estimation from a Scalar Accuracy 21

E.2 Conversion of Reported Performance Measures 23

E.3 Reference Distributions 23

E.4 Normalisation and Uncertainty 24

E.5 Large-Vocabulary Invasive Systems 24

E.6 Table of Estimated Results 24

F Vocabulary Optimisation Experiment Details **24**

A TABLE OF VARIABLES

Table 1 summarises the core variables introduced in Section 2 and 3.

Table 1: Core variables in the open-vocabulary mutual information formulation. Section, proposition, corollary, and equation references point to where each symbol is first introduced.

Symbol	Description	Introduced in
Ω	Reference lexicon.	Section 3
p	Reference distribution over Ω.	Section 3
S	Decoder vocabulary, $S \subset \Omega$.	Section 3
V	Decoder vocabulary size, $V = \lvert S \rvert$.	Section 3
X	User's intended word drawn from the reference distribution, $X \sim p$.	Section 3
Y	Decoder output, $Y \in S \cup \{\varnothing\}$; $Y = \varnothing$ on abstention.	Section 3
Z	In-vocabulary indicator, $Z = \mathbf{1}[X \in S] = \mathbf{1}[Y \neq \varnothing]$.	Section 3
$C(S)$	Lexical coverage of S under p, $C(S) = \Pr(X \in S) = \sum_{w \in S} p(w)$.	Section 3
p_S	In-vocabulary target distribution, $p_S(x) = p(x)/C(S)$ for $x \in S$.	Proposition 2
K_S	In-vocabulary channel, $K_S(y \mid x) = \Pr(Y = y \mid X = x)$ for $x, y \in S$.	Proposition 2
q_S	In-vocabulary output distribution, $q_S(y) = \sum_{x \in S} p_S(x)\, K_S(y \mid x)$.	Proposition 2
P	Per-symbol correct-decoding probability under the homogeneous symmetric channel.	Corollary 1
$h_V(u)$	Entropy of a V-ary symmetric channel row with correct probability u.	Equation 13
$I_{\mathrm{OVMI}}(S)$	Open-vocabulary mutual information, $I_{\mathrm{OVMI}}(S) = C(S)\, I(X; Y \mid X \in S)$.	Def. 1, Eq. 3

B DERIVATION OF WOLPAW'S INFORMATION TRANSFER RATE

Under Wolpaw's standard derivation, two simplifying assumptions are made. First, all intended symbols are equally likely, so $p(x) = 1/V$. Second, the decoder is assumed to have symmetric errors: it is correct with probability P, and when it is wrong, each of the other $V - 1$ symbols is equally likely. Thus

$$\Pr(Y = y \mid X = x) = \begin{cases} P, & y = x, \\ \dfrac{1 - P}{V - 1}, & y \neq x. \end{cases}$$

These assumptions make the calculation straightforward. Fix any output symbol y. There is exactly one intended symbol, namely $x = y$, that produces y as the correct output, contributing probability mass P. The remaining $V - 1$ intended symbols can also produce the same y, but only as an error, and each contributes $(1 - P)/(V - 1)$. Averaging over the uniform prior therefore gives

$$p(y) = \sum_{x \in S} p(y \mid x) p(x) = \frac{1}{V} \left[P + (V - 1) \frac{1 - P}{V - 1} \right] = \frac{1}{V}.$$

So the decoder output Y is itself uniform over the vocabulary, and hence

$$H(Y) = \log_2 V.$$

Next consider the conditional entropy. Once we fix an intended symbol x, the conditional distribution of Y has one outcome with probability P (the correct symbol) and $V-1$ outcomes with probability $(1-P)/(V-1)$ (the incorrect symbols). Therefore

$$H(Y \mid X = x) = -P\log_2 P - (V-1)\frac{1-P}{V-1}\log_2\left(\frac{1-P}{V-1}\right) = -P\log_2 P - (1-P)\log_2\left(\frac{1-P}{V-1}\right).$$

Because this expression does not depend on x, averaging over x leaves it unchanged:

$$H(Y \mid X) = -P\log_2 P - (1-P)\log_2\left(\frac{1-P}{V-1}\right).$$

Substituting these two terms into $I(X;Y) = H(Y) - H(Y \mid X)$ yields the Wolpaw expression for mutual information in an in-vocabulary decoder:

$$I_{\text{Wolpaw}}(X;Y) = \log_2 V + P\log_2 P + (1-P)\log_2\left(\frac{1-P}{V-1}\right). \tag{7}$$

C Proofs and Supplementary Derivations for OVMI

We use the convention $0\log_2 0 = 0$.

C.1 Proof of Proposition 1

Proof of Proposition 1. Recall that

$$Z = \mathbf{1}[X \in S] = \mathbf{1}[Y \neq \varnothing].$$

Since Z is a deterministic function of Y, adjoining Z to the output does not change the mutual information:

$$I(X;Y) = I(X;Y,Z).$$

Applying the chain rule for mutual information gives

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z).$$

Because Z is also a deterministic function of X,

$$I(X;Z) = H(Z) - H(Z \mid X) = H(Z).$$

Now $\Pr(Z=1) = C(S)$ and $\Pr(Z=0) = 1 - C(S)$, so

$$H(Z) = H_2(C(S)).$$

It remains to expand the conditional mutual information:

$$I(X;Y \mid Z) = \sum_{z\in\{0,1\}} \Pr(Z=z)\, I(X;Y \mid Z=z).$$

Under our model, $Z=0$ implies $Y=\varnothing$ deterministically, hence

$$I(X;Y \mid Z=0) = 0.$$

Also, $Z=1$ is exactly the event $X \in S$, so

$$I(X;Y \mid Z=1) = I(X;Y \mid X \in S).$$

Therefore

$$I(X;Y \mid Z) = C(S)\, I(X;Y \mid X \in S).$$

Combining the pieces yields

$$I(X;Y) = H_2(C(S)) + I(X;Y \mid Z) = H_2(C(S)) + C(S)\, I(X;Y \mid X \in S),$$

as claimed. □

C.2 Equivalent Entropy Derivation of Proposition 1

Proposition 1 can also be obtained by decomposing $H(Y)$ and $H(Y \mid X)$.

Because $Z = \mathbf{1}[Y \neq \varnothing]$ is a deterministic function of Y, the chain rule gives

$$H(Y) = H(Z) + H(Y \mid Z).$$

Expanding by the two values of Z,

$$\begin{aligned} H(Y) &= H_2(C(S)) + \Pr(Z=1)\, H(Y \mid Z=1) + \Pr(Z=0)\, H(Y \mid Z=0) \\ &= H_2(C(S)) + C(S)\, H(Y \mid Z=1), \end{aligned} \tag{8}$$

since $Z = 0$ implies $Y = \varnothing$ deterministically and hence $H(Y \mid Z = 0) = 0$. Because $Z = 1$ is equivalent to $X \in S$,

$$H(Y \mid Z=1) = H(Y \mid X \in S).$$

Thus

$$H(Y) = H_2(C(S)) + C(S)\, H(Y \mid X \in S). \tag{9}$$

Similarly,

$$\begin{aligned} H(Y \mid X) &= \Pr(X \in S)\, H(Y \mid X, X \in S) + \Pr(X \notin S)\, H(Y \mid X, X \notin S) \\ &= C(S)\, H(Y \mid X, X \in S), \end{aligned} \tag{10}$$

because $X \notin S$ implies $Y = \varnothing$ deterministically.

Subtracting equation 10 from equation 9,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) \\ &= H_2(C(S)) + C(S)\Big[H(Y \mid X \in S) - H(Y \mid X, X \in S)\Big] \\ &= H_2(C(S)) + C(S)\, I(X;Y \mid X \in S), \end{aligned} \tag{11}$$

recovering equation 2.

C.3 Proposition 2

Proposition 2 (OVMI for a general in-vocabulary decoder)**.** *Now condition on the event that the intended word is representable, i.e. $X \in S$. The corresponding distribution over in-vocabulary target words is $p_S(x) = \Pr(X = x \mid X \in S) = \frac{p(x)}{C(S)}$, where $x \in S$, that is, the unrestricted word distribution restricted to S and renormalised to sum to one.*

For $x, y \in S$, let $K_S(y \mid x) = \Pr(Y = y \mid X = x)$ denote the decoder's conditional distribution over outputs when the intended word is x. Thus K_S is the effective channel operating within the deployed vocabulary.

Under p_S and K_S, the resulting output distribution on S is

$$q_S(y) = \sum_{x \in S} p_S(x)\, K_S(y \mid x), \qquad y \in S.$$

Then

$$I_{\mathrm{OVMI}}(S) = C(S) \left[H(q_S) - \sum_{x \in S} p_S(x)\, H(K_S(\cdot \mid x)) \right], \tag{12}$$

where entropy $H(W) = -\sum_i p(w_i) \log_2 p(w_i)$ for a discrete random variable W.

C.4 Proof of Proposition 2

Proof of Proposition 2. By Definition 1,

$$I_{\mathrm{OVMI}}(S) = C(S)\, I(X;Y \mid X \in S).$$

Conditioned on $X \in S$, the input distribution is

$$p_S(x) = \Pr(X = x \mid X \in S) = \frac{p(x)}{C(S)},$$

and the in-vocabulary channel is

$$K_S(y \mid x) = \Pr(Y = y \mid X = x), \qquad x, y \in S.$$

The output distribution is therefore

$$q_S(y) = \Pr(Y = y \mid X \in S) = \sum_{x \in S} p_S(x)\, K_S(y \mid x).$$

Applying the standard mutual-information identity $I(X;Y) = H(Y) - H(Y \mid X)$ under the conditional distribution $X \mid X \in S$,

$$I(X;Y \mid X \in S) = H(q_S) - \sum_{x \in S} p_S(x)\, H(K_S(\cdot \mid x))\,.$$

Multiplying by $C(S)$ yields

$$I_{\mathrm{OVMI}}(S) = C(S) \left[H(q_S) - \sum_{x \in S} p_S(x)\, H(K_S(\cdot \mid x)) \right],$$

which is equation 12. □

C.5 Proof of Corollary 1

We will use the following shorthand

$$h_V(u) := -u \log_2 u - (1-u) \log_2 \left(\frac{1-u}{V-1} \right). \tag{13}$$

This is the same as the term $H(Y|X)$ in Wolpaw's formula in Eq. 1, describing the entropy of the decoder's output distribution under a symmetric error assumption where the intended word is output with probability u, and each of the other $V-1$ words is output with probability $(1-u)/(V-1)$.

Proof of Corollary 1. Under the homogeneous symmetric channel,

$$K_S(y \mid x) = \begin{cases} P, & y = x, \\ \dfrac{1-P}{V-1}, & y \neq x. \end{cases}$$

For each $x \in S$, the entropy of the corresponding channel row is

$$H(K_S(\cdot \mid x)) = -P \log_2 P - (1-P) \log_2 \left(\frac{1-P}{V-1} \right) = h_V(P),$$

where h_V is defined in equation 13.

The induced output distribution satisfies

$$\begin{aligned} q_S(j) &= \sum_{x \in S} p_S(x)\, K_S(j \mid x) \\ &= p_S(j) P + \sum_{x \neq j} p_S(x) \frac{1-P}{V-1} \\ &= P\, p_S(j) + \frac{1-P}{V-1} \big(1 - p_S(j)\big), \end{aligned} \tag{14}$$

which is equation 4.

Substituting into Proposition 2 gives

$$\begin{aligned} \widehat{I}^{\,\mathrm{scalar}}_{\mathrm{OVMI}}(S) &= C(S) \left[H(q_S) - \sum_{x \in S} p_S(x)\, h_V(P) \right] \\ &= C(S) \Big[H(q_S) - h_V(P) \Big], \end{aligned} \tag{15}$$

because $\sum_{x \in S} p_S(x) = 1$. This is equation 5. □

C.6 Retaining Per-Word Accuracies in OVMI

Corollary 2 (Symmetric error model with word-specific accuracies)**.** *Suppose each intended word $x \in S$ has its own correct-decoding probability $P_c(x)$. Thus some words may be easier to decode than others. Conditional on an error, however, the decoder distributes the remaining probability uniformly across the other $V-1$ in-vocabulary words. Hence*

$$K_S(y \mid x) = \begin{cases} P_c(x), & y = x, \\ \dfrac{1 - P_c(x)}{V-1}, & y \neq x. \end{cases}$$

Under this model, the output distribution on S is

$$q_S(j) = p_S(j)P_c(j) + \frac{1}{V-1}\sum_{x \neq j} p_S(x)\big(1 - P_c(x)\big), \qquad j \in S. \tag{16}$$

The first term is the probability mass from correctly decoding j when $X = j$. The second term is the total probability mass assigned to j when some other in-vocabulary word was intended but the decoder makes an error and spreads that error uniformly across the incorrect outputs. Consequently,

$$\widehat{I}^{\text{word}}_{\text{OVMI}}(S) = C(S)\left[H(q_S) - \sum_{x \in S} p_S(x)\, h_V\big(P_c(x)\big)\right]. \tag{17}$$

This model relaxes the shared-accuracy assumption of Corollary 1, while retaining the simplifying assumption that, conditional on an error, all incorrect outputs are equally likely.

Proof of Corollary 2. Under the word-dependent symmetric channel,

$$K_S(y \mid x) = \begin{cases} P_c(x), & y = x, \\ \dfrac{1 - P_c(x)}{V-1}, & y \neq x. \end{cases}$$

For each $x \in S$, the entropy of the x-th channel row is

$$H(K_S(\cdot \mid x)) = -P_c(x)\log_2 P_c(x) - (1 - P_c(x))\log_2\left(\frac{1 - P_c(x)}{V-1}\right) = h_V\big(P_c(x)\big).$$

The induced output distribution is

$$\begin{aligned} q_S(j) &= \sum_{x \in S} p_S(x)\, K_S(j \mid x) \\ &= p_S(j)P_c(j) + \sum_{x \neq j} p_S(x)\frac{1 - P_c(x)}{V-1} \\ &= p_S(j)P_c(j) + \frac{1}{V-1}\sum_{x \neq j} p_S(x)\big(1 - P_c(x)\big), \end{aligned} \tag{18}$$

which is equation 16.

Substituting into Proposition 2 yields

$$\widehat{I}^{\text{word}}_{\text{OVMI}}(S) = C(S)\left[H(q_S) - \sum_{x \in S} p_S(x)\, h_V\big(P_c(x)\big)\right],$$

which is equation 17. □

C.7 Further Observations for the Homogeneous Symmetric Channel

A useful rearrangement of equation 4 is

$$q_S(j) = \frac{1-P}{V-1} + \frac{PV-1}{V-1}\, p_S(j). \tag{19}$$

Equation 19 makes three facts immediate.

First, if p_S is uniform, then q_S is uniform.

Second, if $P = 1/V$ (chance performance), then q_S is uniform for any p_S.

Third, if $P \neq 1/V$, then q_S is non-uniform if and only if p_S is non-uniform. In particular, for informative decoders with $P > 1/V$, any non-uniformity in the in-vocabulary language distribution propagates to the output distribution. Consequently,

$$H(q_S) < \log_2 V$$

whenever p_S is non-uniform and $P \neq 1/V$, since the uniform distribution uniquely maximises entropy on a finite alphabet.

D Further Notes on OVMI

D.1 Relation to Continuous Sequence Decoding

The abstention model used in Proposition 1 is an idealised word-level model in which knowing whether the decoder supports a word is observable from the decoder output: if $X \notin S$, the user abstains and $Y = \varnothing$, whereas if $X \in S$, the decoder produces an output in S.

Continuous speech decoders, such as generative models that autoregressively produce sub-word tokens, need not satisfy this assumption exactly. An in-vocabulary word may be deleted, producing no aligned output despite $X \in S$, while an out-of-vocabulary intended word may instead result in substitution with an in-vocabulary word. Insertions have no corresponding intended word at all. Consequently, Proposition 1 should be interpreted as a decomposition of the output word sequence.

For retrospective studies that report only word error rate, we therefore use $P = 1 - \mathrm{WER}$ as a scalar proxy for in-vocabulary decoding fidelity and apply the estimator of Corollary 1. This collapses substitutions, deletions, and insertions into a single error statistic. In particular,

$$1 - \mathrm{WER} = 1 - \frac{N_{\mathrm{sub}} + N_{\mathrm{del}} + N_{\mathrm{ins}}}{N} \leq 1 - \frac{N_{\mathrm{sub}} + N_{\mathrm{del}}}{N},$$

so insertions make this quantity a conservative proxy for the fraction of reference words recovered. When word-aligned predictions or richer sequence-level statistics are available, these errors should instead be modelled explicitly with the per-word or general estimators we define.

D.2 Reference Reweighting and Channel Invariance

Let d denote the domain in which decoding capability is measured and write

$$K_{S,d}(y \mid x) = \Pr(Y = y \mid X = x, d).$$

Given a channel measured in domain d, evaluating it against another reference distribution p' computes

$$I_{\mathrm{OVMI}}(p', K_{S,d}).$$

This changes the weighting of intended words while retaining the measured conditional decoder channel. This quantity is distinct from the OVMI that would be measured if the decoder were re-evaluated in another domain d', $I_{\mathrm{OVMI}}(p', K_{S,d'})$. The two coincide when

$$K_{S,d}(y \mid x) = K_{S,d'}(y \mid x)$$

for all relevant x, y, which we call the *channel invariance assumption*. This assumption is fairly natural for isolated word decoding, but may only be approximate for continuous decoders whose errors depend on linguistic context, speaker or task distribution, or language-model priors. OVMI should be interpreted as reweighting a measured decoder's capability, which is not necessarily the same as predicting performance in another domain.

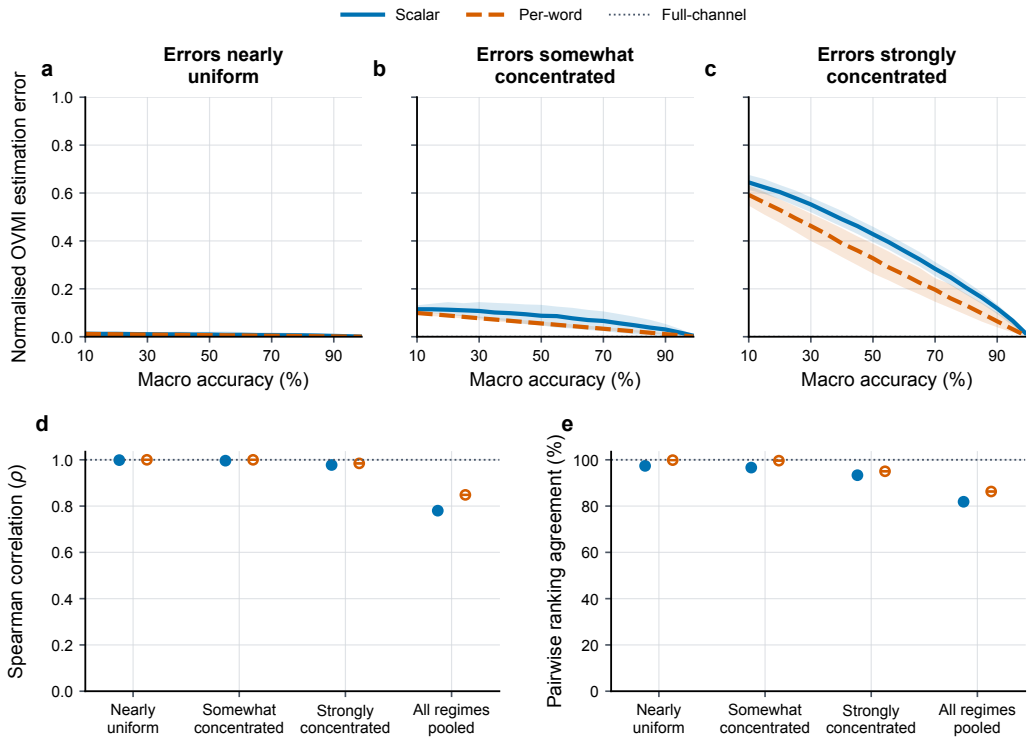

Figure 7: **OVMI estimators compared with full-channel OVMI for simulated 50-word decoders. (a–c)** Normalised estimation error, $|I_{\mathrm{OVMI}}^{\mathrm{estimate}} - I_{\mathrm{OVMI}}^{\mathrm{full}}|/I_{\mathrm{OVMI}}^{\max}$, where $I_{\mathrm{OVMI}}^{\max} = C(S)H(p_S) = 2.360$ bits for the top-50 SUBTLEX-UK vocabulary. Curves show medians and shading shows the 5th–95th percentiles. **(d)** Spearman correlation and **(e)** pairwise ranking agreement with full-channel OVMI; error bars show 95% intervals from 1,000 bootstrap resamples. Error concentration is $c = 1 - \frac{1}{50}\sum_{i=1}^{50} H(R_i)/\log_2(49)$, where R_i is the distribution over incorrect outputs conditional on intended word i being decoded incorrectly. Thus, $c = 0$ indicates uniform errors over the 49 incorrect outputs and $c = 1$ indicates that errors for each word are concentrated on a single incorrect output. Panels a–c have $c \approx 0.01, 0.11, 0.69$, respectively.

D.3 Robustness of OVMI Estimators

Our retrospective comparisons often require estimating OVMI from less information than the full decoder channel. We therefore test how the scalar and per-word estimators behave as the true channel departs from their assumptions (Figure 7). We simulate decoders over the 50 most frequent words in SUBTLEX-UK while holding the reference distribution fixed, and vary both macro accuracy and channel structure. Specifically, we consider increasingly heterogeneous per-word accuracies and increasingly concentrated off-diagonal errors, ranging from errors distributed approximately uniformly among incorrect words to errors concentrated on a small number of outputs. The scalar estimator represents all words with a single macro accuracy and assumes uniform off-diagonal errors; the per-word estimator captures differences in word-level accuracy but retains the uniform off-diagonal assumption. The general (full-channel) estimator uses the complete simulated confusion matrix and is our reference.

Experimental setup. We simulate 19 target macro accuracies, $P \in \{0.10, 0.15, \ldots, 0.95, 0.99\}$, generating 600 channels per accuracy and error regime for a total of 34,200 channels. For each channel, word-specific accuracies are generated as $a_i = \mathrm{sigmoid}(b + \sigma z_i)$, where the z_i are standardised Gaussian draws and b is chosen numerically so that the mean of the 50 accuracies is exactly P. Conditional on an error for word i, its probability mass is distributed among the other 49 words according to an independently sampled $R_i \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha)$, giving a confusion matrix Q where $Q_{ii} = a_i$

and $Q_{ij} = (1 - a_i)R_{ij}$ for $j \neq i$. The three regimes use $(\sigma, \alpha) = (0.3, 10), (1.5, 1), (6, 0.03)$, jointly increasing heterogeneity in per-word accuracy and concentration of off-diagonal errors. Full-channel OVMI is computed directly from the resulting 50×50 channel Q, while the per-word approximation retains only the a_i and the scalar approximation retains only their mean P, with both approximations replacing the observed off-diagonal structure by uniform errors.

Results. Figures 7a–c show that both approximations closely recover full-channel OVMI when errors are approximately uniform. Approximation error increases as errors become concentrated, particularly at lower decoding accuracy. This follows from the channel assumptions where for a fixed error probability, distributing errors uniformly across incorrect outputs produces greater conditional entropy than concentrating them on a small number of systematic confusions. The scalar and per-word estimators therefore cannot capture information contained in the structure of incorrect predictions. Per-word estimates are consistently more accurate because they additionally retain heterogeneity in correct-decoding probabilities.

For retrospective comparison, we are also interested in whether an approximation preserves the ordering of systems even when its absolute OVMI values are biased. We quantify this using Spearman rank correlation and pairwise ranking agreement, assigning half credit to tied estimates (Figures 7d–e). Within each error regime, ordering is highly stable as even under strongly concentrated errors, scalar OVMI achieves a Spearman correlation of approximately 0.98 and pairwise agreement above 90%, with the per-word estimator performing better still. Agreement decreases when channels from all error regimes are pooled because the magnitude of approximation bias depends on channel structure. Systems with substantially different error structures can therefore be shifted by different amounts.

These results distinguish the roles of the three estimators. Full-channel OVMI should be preferred whenever a reliable confusion matrix is available, while per-word accuracies provide a stronger approximation when the full error structure is unavailable. The scalar estimator enables retrospective evaluation of studies that report only aggregate accuracy or WER and generally preserves comparative ordering within similar channel regimes. However, its absolute value should not be interpreted as a substitute for full-channel OVMI when decoding errors are strongly concentrated.

D.4 Closed Form OVMI Pseudocode

In Figure 8, we provide Python pseudocode for computing OVMI under the homogeneous (single accuracy value) and symmetric error assumption. Note that when computing OVMI, we typically exclude words with fewer than five instances to ensure accuracy estimates for P_{macro} are reliable.

E Cross-Study OVMI Estimation Details

This appendix describes how we estimate OVMI from the summary statistics reported by existing speech-decoding studies and how we construct the reference distributions used in Section 5. The general definition of OVMI in Proposition 2 requires the decoder channel $K_S(y \mid x)$. In retrospective comparisons, however, published studies rarely report complete confusion matrices and instead typically provide a vocabulary size together with accuracy or word error rate (WER). We therefore use the homogeneous symmetric-channel estimator from Eq. 5 whenever a more detailed channel estimate is unavailable.

E.1 Estimation from a Scalar Accuracy

For a decoder vocabulary S of size V, and given a single correct-decoding probability P, we approximate the in-vocabulary channel as

$$K_S(y \mid x) = \begin{cases} P, & y = x, \\ \dfrac{1 - P}{V - 1}, & y \neq x. \end{cases}$$

The induced output distribution is therefore

$$q_S(y) = P p_S(y) + \frac{1 - P}{V - 1} \left[1 - p_S(y)\right],$$

```python
import numpy as np
from scipy.special import xlogy

_LN2 = np.log(2.0)

def _xlog2y(x, y):
    """x * log2(y), with the convention 0 * log2(0) = 0."""
    return xlogy(x, y) / _LN2

def compute_ovmi(subset, P_c, freqs, p_freq):
    """
    subset: indices of words in the candidate vocabulary
    P_c:    macro-averaged classification accuracy over the subset
    freqs:  reference unigram frequencies (aligned with subset)
    p_freq: total number of words in the reference distribution
    Returns the open-vocabulary mutual information (OVMI).
    """
    V = len(subset)
    if V <= 1:
        return 0.0

    # (1) Coverage: natural-language mass spanned by the subset
    coverage = freqs.sum() / p_freq

    # (2) Input distribution p_s(x) over subset words
    if freqs.sum() != 0:
        p_x = freqs / freqs.sum()
    else:
        return 0.0

    # (3) Output distribution q(y) assuming errors spread uniformly
    #     over the V-1 incorrect classes
    p_err = (1 - P_c) / (V - 1)
    q_y = p_err + p_x * (P_c - p_err)

    # (4) Marginal entropy of predictions
    H_Y = -np.sum(_xlog2y(q_y, q_y))

    # (5) Conditional entropy
    H_YX = -_xlog2y(P_c, P_c) - _xlog2y(1 - P_c, p_err)

    # (6) Mutual information, floored at 0 to absorb numerical noise
    MI = max(0.0, H_Y - H_YX)

    # (7) Weight by natural-language coverage
    return coverage * MI
```

Figure 8: OVMI computation for a candidate vocabulary subset with a single accuracy scalar.

and the conditional entropy of the symmetric channel is

$$h_V(P) = -P \log_2 P - (1-P) \log_2 \frac{1-P}{V-1}.$$

The retrospective estimate used throughout the cross-study comparison is then

$$\widehat{I}_{\mathrm{OVMI}} = C(S)\,[H(q_S) - h_V(P)]\,.$$

This approximation assumes only that all supported words have the same correct-decoding probability and that errors are distributed uniformly among the remaining $V-1$ outputs. The reference distribution itself remains non-uniform through $C(S)$, p_S, and q_S.

Where available, we use macro rather than micro accuracy for P. In particular,

$$P_{\text{macro}} = \frac{1}{V} \sum_{x \in S} \Pr(\hat{Y} = x \mid X = x),$$

which summarises the reliability of an average supported word without additionally weighting common words by their frequency. Frequency is already represented by the external distribution p.

E.2 Conversion of Reported Performance Measures

Balanced top-1 accuracy. For the non-invasive word-classification results, the reported balanced top-1 accuracy is the macro-average of the class-wise recalls.

Isolated-word accuracy. For studies reporting accuracy on an isolated V-way word-classification task, we use the reported mean correct-classification probability directly as P. This gives $P = 0.471$ for the isolated 50-word experiment of Moses et al. (2021) and $P = 0.940$ for the corresponding 50-word result of Willett et al. (2023).

Word error rate. Continuous speech decoders generally report WER rather than word classification accuracy. For a reference containing N words, with N_{sub} substitutions, N_{del} deletions, and N_{ins} insertions,

$$\text{WER} = \frac{N_{\text{sub}} + N_{\text{del}} + N_{\text{ins}}}{N}.$$

The proportion of intended words that are neither substituted nor deleted is

$$P_{\text{correct}} = 1 - \frac{N_{\text{sub}} + N_{\text{del}}}{N} = 1 - \text{WER} + \frac{N_{\text{ins}}}{N}.$$

Hence

$$P_{\text{correct}} \geq 1 - \text{WER}.$$

When only WER is available, we therefore set

$$P = 1 - \text{WER}.$$

This is a conservative estimate of per-intended-word correctness because insertions increase WER without corresponding to a failure to recover an intended word. For the language-model-assisted results, this gives $P = 0.744$ for Moses, $P = 0.909$ and 0.762 for the 50-word and 125k-word Willett systems, respectively, and $P = 0.975$ for Card.

E.3 Reference Distributions

All reference distributions are unigram distributions over word types. For a corpus-derived reference, we estimate

$$p(w) = \frac{n(w)}{\sum_{u \in \Omega} n(u)},$$

where $n(w)$ is the number of occurrences of word w in the reference corpus after lower-casing all words. For a decoder vocabulary S, words outside S remain part of the reference distribution and contribute to the uncovered probability mass $1 - C(S)$; p is therefore never renormalised to the decoder vocabulary before computing coverage.

Broad spoken English: SUBTLEX-UK. Our default reference is SUBTLEX-UK (van Heuven et al., 2014), a word-frequency norm derived from British film and television subtitles. We normalise the SUBTLEX-UK frequency counts to obtain p. This distribution has entropy $H(p) = 9.77$ bits after the preprocessing used in our evaluation and is intended as a broad spoken-English reference.

Conversation: Switchboard. For conversational speech, we construct an empirical unigram distribution from the 36-call Switchboard sample distributed with NLTK. Word counts are aggregated across the sample and normalised to sum to one. The resulting reference has entropy $H(p) = 8.27$ bits.

AAC: Universal Core Vocabulary. For an assistive communication reference, we use the 36-word Universal Core Vocabulary (UCV) of Erickson et al. (2019). UCV specifies a set of words rather than a probability distribution. We therefore assign each UCV word its corresponding SUBTLEX-UK frequency and renormalise within the UCV set,

$$p_{\mathrm{UCV}}(w) = \frac{f_{\mathrm{SUBTLEX}}(w)}{\sum_{u \in \mathrm{UCV}} f_{\mathrm{SUBTLEX}}(u)}, \qquad w \in \mathrm{UCV}.$$

This preserves natural differences in word frequency rather than treating the 36 words as equiprobable. The resulting distribution has $H(p) = 4.30$ bits.

Narrative speech: Sherlock. For narrative prose, we use the held-out LibriBrain100 text corresponding to Session 12, Chapter 1 of *The Adventures of Sherlock Holmes* (Doyle, 1892). This session belongs to the LibriBrain100 test set and is not used to fit the non-invasive decoders. We count the words occurring in the chapter and normalise their empirical frequencies to form the narrative reference distribution. The resulting distribution has entropy $H(p) = 8.44$ bits. This reference is intentionally narrower than SUBTLEX-UK and represents the narrative-speech domain on which several of the non-invasive systems are evaluated.

E.4 Normalisation and Uncertainty

OVMI is reported in bits per intended word drawn from p. When comparing performance across different reference distributions, we additionally normalise by the entropy of that distribution,

$$\mathrm{OVMI}_{\%} = 100\,\frac{I_{\mathrm{OVMI}}}{H(p)}.$$

This quantity is the percentage of lexical information available under the stated reference distribution that is conveyed by the system.

For invasive isolated-word results, uncertainty in the reported accuracy is propagated through the OVMI estimator using Wilson score intervals. For WER-derived results, we propagate the corresponding published sentence- or trial-level uncertainty through $P = 1 - \mathrm{WER}$. For the non-invasive experiments, we compute OVMI at the mean balanced accuracy and at one standard error above and below that mean across training seeds.

E.5 Large-Vocabulary Invasive Systems

The 125k-word decoder lexicons used by Willett et al. (2023) and Card et al. (2024) come from a lexicon based on the CMU Pronouncing Dictionary, but are not published in full as far as we are aware. For coverage calculations, we use the CMU Pronouncing Dictionary ourselves, after collapsing pronunciation variants to unique orthographic word forms. This yields 126,052 unique spellings. We retain the reported vocabulary size $V = 125{,}000$ when evaluating the decoder channel.

E.6 Table of Estimated Results

Table 2 provides OVMI scores and uncertainties for the systems discussed in the main text across all the reference distributions explored in this work. We use the scalar approximation (Corollary 1) with the WER proxy when only WER is quoted by the study (typically in invasive systems).

F Vocabulary Optimisation Experiment Details

We compare OVMI-based vocabulary selection against additional baselines that maximise validation accuracy, or use random selection in Figure 9. The figure also shows absolute accuracy across methods, including for frequency selection. Accuracy is computed according to the ground-truth sentence, where anything out-of-vocabulary is treated as an error.

Data and splits. We use subject 0 from LibriBrain100 and report results for TIMIT, Podcasts, and Sherlock. Reference word distributions are estimated exclusively from the training portion of the

Table 2: **Retrospective cross-study OVMI estimates.** Each cell reports OVMI in bits, with OVMI/$H(p)$ in parentheses and uncertainty beneath. Rows are ordered within each block by SUBTLEX–UK OVMI. Superscripts on P identify balanced top-1 accuracy (b), isolated-word accuracy (a), and $1 - \mathrm{WER}$ (w). Bracketed ranges are 95% sampling intervals for invasive systems using Wilson score intervals for isolated-word accuracies and published sentence/trial intervals for WER-derived rows. $\pm$ values are OVMI uncertainties obtained by mapping mean P plus or minus one SEM across three training seeds for non-invasive systems.

System	V	P	Broad spoken SUBTLEX–UK $H(p) = 9.77$	Conversational Switchboard $H(p) = 8.27$	AAC / clinical UCV $H(p) = 4.30$	Narrative prose Sherlock $H(p) = 8.44$
Attempted speech (invasive)						
Card (+LM)	125k	97.5%^{w}	9.156 (93.7%) [9.093, 9.208]	8.015 (96.9%) [7.964, 8.057]	4.197 (97.5%) [4.171, 4.218]	8.096 (95.9%) [8.046, 8.138]
Willett (+LM)	125k	76.2%^{w}	7.038 (72.0%) [6.835, 7.232]	6.227 (75.3%) [6.052, 6.394]	3.279 (76.2%) [3.189, 3.365]	6.311 (74.7%) [6.135, 6.479]
Metzger (+LM)	1024	74.5%^{w}	3.712 (38.0%) [3.184, 4.089]	4.028 (48.7%) [3.462, 4.430]	3.148 (73.1%) [2.747, 3.426]	3.327 (39.4%) [2.866, 3.653]
Willett (isolated)	50	94.0%^{a}	0.656 (6.7%) [0.637, 0.672]	0.949 (11.5%) [0.922, 0.971]	1.822 (42.3%) [1.776, 1.859]	0.239 (2.8%) [0.232, 0.244]
Willett (+LM)	50	90.9%^{w}	0.621 (6.4%) [0.599, 0.642]	0.900 (10.9%) [0.868, 0.929]	1.737 (40.4%) [1.681, 1.789]	0.227 (2.7%) [0.219, 0.234]
Moses (+LM)	50	74.4%^{w}	0.459 (4.7%) [0.360, 0.539]	0.670 (8.1%) [0.527, 0.783]	1.321 (30.7%) [1.053, 1.529]	0.169 (2.0%) [0.133, 0.197]
Moses (isolated)	50	47.1%^{a}	0.238 (2.4%) [0.230, 0.245]	0.350 (4.2%) [0.339, 0.361]	0.711 (16.5%) [0.690, 0.733]	0.088 (1.0%) [0.085, 0.091]
Perceived speech (non-invasive)						
Tang 2023	6867	6.7%^{w}	0.357 (3.6%) $\pm$0.030	0.376 (4.5%) $\pm$0.031	0.266 (6.2%) $\pm$0.020	0.348 (4.1%) $\pm$0.029
LibriBrain100 2025	50	25.8%^{b}	0.237 (2.4%) $\pm$0.003	0.275 (3.3%) $\pm$0.003	0.316 (7.3%) $\pm$0.004	0.221 (2.6%) $\pm$0.003
Armeni 2022	50	20.8%^{b}	0.176 (1.8%) $\pm$0.004	0.192 (2.3%) $\pm$0.004	0.254 (5.9%) $\pm$0.005	0.189 (2.2%) $\pm$0.004
MEG-MASC 2023	50	8.5%^{b}	0.033 (0.3%) $\pm$0.003	0.038 (0.5%) $\pm$0.004	0.059 (1.4%) $\pm$0.006	0.035 (0.4%) $\pm$0.003

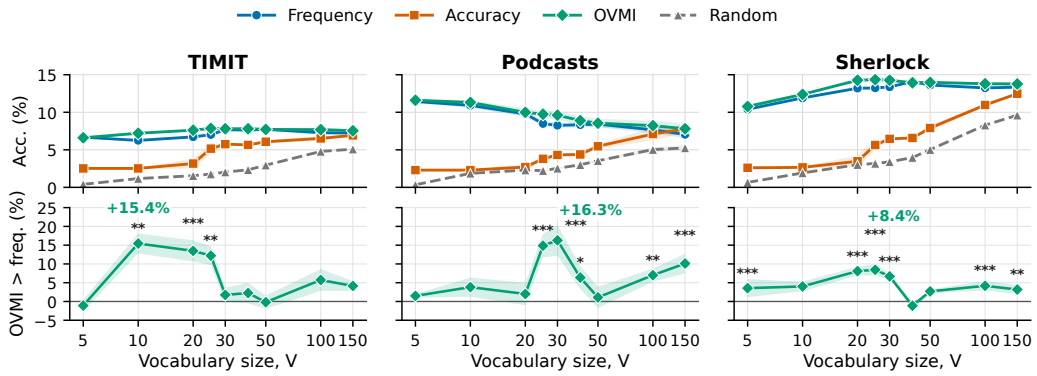

Figure 9: **OVMI-guided vocabulary selection improves accuracy.** We compare frequency, validation accuracy, OVMI, and random selection across three speech domains. Top: held-out accuracy. Bottom: relative improvement of OVMI over frequency. Shading denotes SEM over five seeds and stars paired permutation-test significance ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$).

corresponding domain. For Podcasts, sessions 1–28 form the training set, session 29 is validation, and session 30 is test. For Sherlock, the training set comprises the training runs, with `Sherlock1` session 11 used for validation and session 12 for test. TIMIT is split at the utterance level using the canonical LibriBrain100 policy: the 50 development speakers are used for validation and the 24 core-test speakers for testing. The repeated SA utterances are excluded from validation and test, as are training speakers sharing an SX sentence with a core-test speaker. MOCHA-TIMIT training sessions A and D contribute only to decoder training and candidate-word availability.

Candidate vocabulary and neural decoder. The candidate pool is constructed from word tokens with usable neural training windows after removal of incomplete examples. Words must occur at least five times in the pooled neural training data. Eligible words are ordered by decreasing training count and the pool is capped at 250 words. The same candidate pool is used for all domains, selection methods, vocabulary sizes, and model seeds.

We train five independently initialised instances of the contrastive brain-to-T5-embedding decoder by d'Ascoli et al. (2025) for 50 epochs with a batch size of 128. For each seed, the checkpoint is selected using pooled validation contrastive top-1 accuracy. The model has no vocabulary-specific classification head. After training, we construct one normalised target embedding for each candidate word and cache the cosine score between every validation or test neural prediction and all 250 candidates. Consequently, evaluating a vocabulary requires only restricting the cached score matrix to the corresponding columns and applying an argmax; the neural model is never retrained for a particular method or vocabulary size.

Reference distribution. For domain d, let $p_d(w)$ denote the empirical probability of word w in that domain's training text. Probabilities are defined over the full training-text distribution rather than being renormalised over the 250-word candidate pool. Thus, for a selected vocabulary $\mathcal{V}$, its lexical coverage is

$$C_d(\mathcal{V}) = \sum_{w \in \mathcal{V}} p_d(w).$$

No validation/test text contributes to p_d. We then use the full confusion-channel estimate for OVMI.

Greedy optimisation. Starting from an empty vocabulary, we evaluate each remaining candidate and add the word that gives the highest validation-estimated OVMI for the resulting set. We repeat until the vocabulary contains V words. Since OVMI is zero for every singleton, the first word maximises $p_d(w)\hat{P}_c(w)$, where $\hat{P}_c(w)$ is its estimated validation decoding accuracy. Ties favour higher training-text probability, then higher validation decoding accuracy, then alphabetical order.

Statistical testing. For every domain and vocabulary size, OVMI and Frequency are compared using a one-sided paired randomisation test. We use $B = 100{,}000$ Monte Carlo permutations and calculate the one-sided p-value with the finite-simulation correction

$$p_{\text{raw}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}[T_b \geq T_{\text{obs}}]}{B + 1}.$$

Each V is tested separately. Holm's step-down correction is then applied across the nine vocabulary sizes within each domain, controlling the error rate separately for TIMIT, Podcasts, and Sherlock.