



OPEN ACCESS

EDITED BY

Jordi Solé-Casals,
University of Vic - Central University of
Catalonia, Spain

REVIEWED BY

Guilherme Wood,
University of Graz, Austria
Kamil A. Grajski,
NuroSci, LLC, United States
Anarghya Das,
University at Buffalo, United States

*CORRESPONDENCE

Esra Sümer-Arpak
✉ esra.sumer.arpak@associated.ltu.se

RECEIVED 24 March 2026

REVISED 01 May 2026

ACCEPTED 05 May 2026

PUBLISHED 28 May 2026

CITATION

Sümer-Arpak E, Saini R, Chakladar DD,
Varun SK and Simistira Liwicki F (2026)
The current status of foundation models
in decoding inner speech from
non-invasive brain signals: a mini review.
Front. Hum. Neurosci. 20:1838064.
doi: 10.3389/fnhum.2026.1838064

COPYRIGHT

© 2026 Sümer-Arpak, Saini, Chakladar,
Varun and Simistira Liwicki. This is an
open-access article distributed under the
terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

The current status of foundation models in decoding inner speech from non-invasive brain signals: a mini review

Esra Sümer-Arpak*, Rajkumar Saini, Debashis Das Chakladar,
Sanjeev Kumar Varun and Foteini Simistira Liwicki

Division of Embedded Intelligent Systems LAB, Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden

Inner speech (IS), or imagined speech without overt articulation, is a promising target for brain-computer interfaces (BCIs) aimed at restoring communication in individuals with severe speech impairments, such as locked-in syndrome. Foundation models (FMs), typically trained using self-supervised learning (SSL) on large-scale datasets, offer new opportunities for learning transferable and robust representations from neural signals. This mini review provides an overview of FM-based approaches for IS decoding using non-invasive neuroimaging modalities, including functional magnetic resonance imaging, electroencephalography, magnetoencephalography, and functional near-infrared spectroscopy, highlighting architectural trends, pretraining strategies, and model adaptation techniques. We discuss how recent models move beyond task-specific classification toward scalable representation learning and semantic-level decoding. Despite these advances, several challenges remain, including the weak, noisy, and non-stationary nature of neural signals, variability in data acquisition, and limitations in dataset scale, standardization, computational resources, interpretability, and evaluation metrics. Ethical and privacy considerations are also critical. Overall, FMs provide a promising paradigm for non-invasive IS decoding, addressing neurophysiological, methodological, and ethical challenges is essential for developing scalable and reliable BCI systems.

KEYWORDS

deep learning, foundation models, inner speech decoding, neural signals, non-invasive neuro imaging

1 Introduction

Speech is fundamental to daily communication, yet neurological disorders, trauma, and disease can impair this ability (Shah et al., 2022). Inner speech (IS), also referred to as verbal thinking or covert self-talk, describes the internal experience of language without overt or subvocal articulation (Alderson-Day and Fernyhough, 2015). IS encompasses related paradigms, including silent or intended speech (attempted articulation without sound), articulatory motor imagery (imagined speech), and phonological rehearsal (the internal repetition of speech sounds to support working memory) (Nieto et al., 2022; Schultz et al., 2017). These paradigms involve different neural systems and signal characteristics, affecting decoding performance. In this review, we adopt Vygotsky's model (Vygotsky, 1987), defining IS

as an internalized process of thinking in pure meanings, distinct from motor imagery or phonological rehearsal (Schultz et al., 2017). Decoding IS from non-invasive brain signals holds promise for brain–computer interfaces (BCIs) aimed at restoring communication in individuals with severe impairments, such as those with locked-in syndrome (LIS).

IS elicits activity in speech-related brain regions (Shergill et al., 2001). Its neural basis has been investigated using multiple neuroimaging modalities, including functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG), functional near-infrared spectroscopy (fNIRS), and electrocorticography (ECoG). Among them, ECoG provides high-resolution intracranial recordings of neural activity (Martin et al., 2018). Several studies have exploited these advantages to achieve higher decoding performance (Pei et al., 2011; Martin et al., 2016; Komeiji et al., 2024). However, it requires invasive electrodes, which limits its scalability and generalizability (Martin et al., 2018). In contrast, non-invasive modalities, despite typically yielding lower performance, offer scalable solutions and have demonstrated promising results for IS decoding (Almufareh et al., 2025). Detailed comparisons of these modalities, including their signal principles, spatial–temporal resolution, and associated trade-offs, are summarized in [Supplementary Table S1](#). These differences affect decoding performance and motivate multimodal approaches and advanced analytical methods for reliable IS decoding (Cooney et al., 2022; Wellington et al., 2024).

Recent advancements in machine learning and computational power have strengthened links between cognitive neuroscience and practical applications, such as decoding IS from brain signals. IS decoding can be performed using traditional machine learning classifiers or deep learning models. Classical techniques such as support vector machines (SVM) (Wellington et al., 2024), Random forests (Hernández-Del-Toro et al., 2021), and gradient boosting machines (Pan et al., 2023) have been used for feature-based classification of IS. Conversely, deep learning models do not require manual feature extraction. Supervised deep learning models, such as convolutional neural networks (CNNs) (Simistira Liwicki et al., 2022), recurrent neural networks (RNNs) (Chengaiyan et al., 2020), and hybrid methods (Saha and Fels, 2019), have achieved remarkable results. However, supervised deep learning models need large labeled datasets, which limits the robustness and generalization of these models (Bommasani et al., 2021). Recently proposed foundational models (FMs) have gained significant attention because they are trained on pretext tasks with unlabeled data and can then be applied to multiple downstream tasks (Gui et al., 2024). These approaches have already proven successful in learning robust representations from brain signals and are increasingly explored for IS decoding (Lesaja et al., 2022).

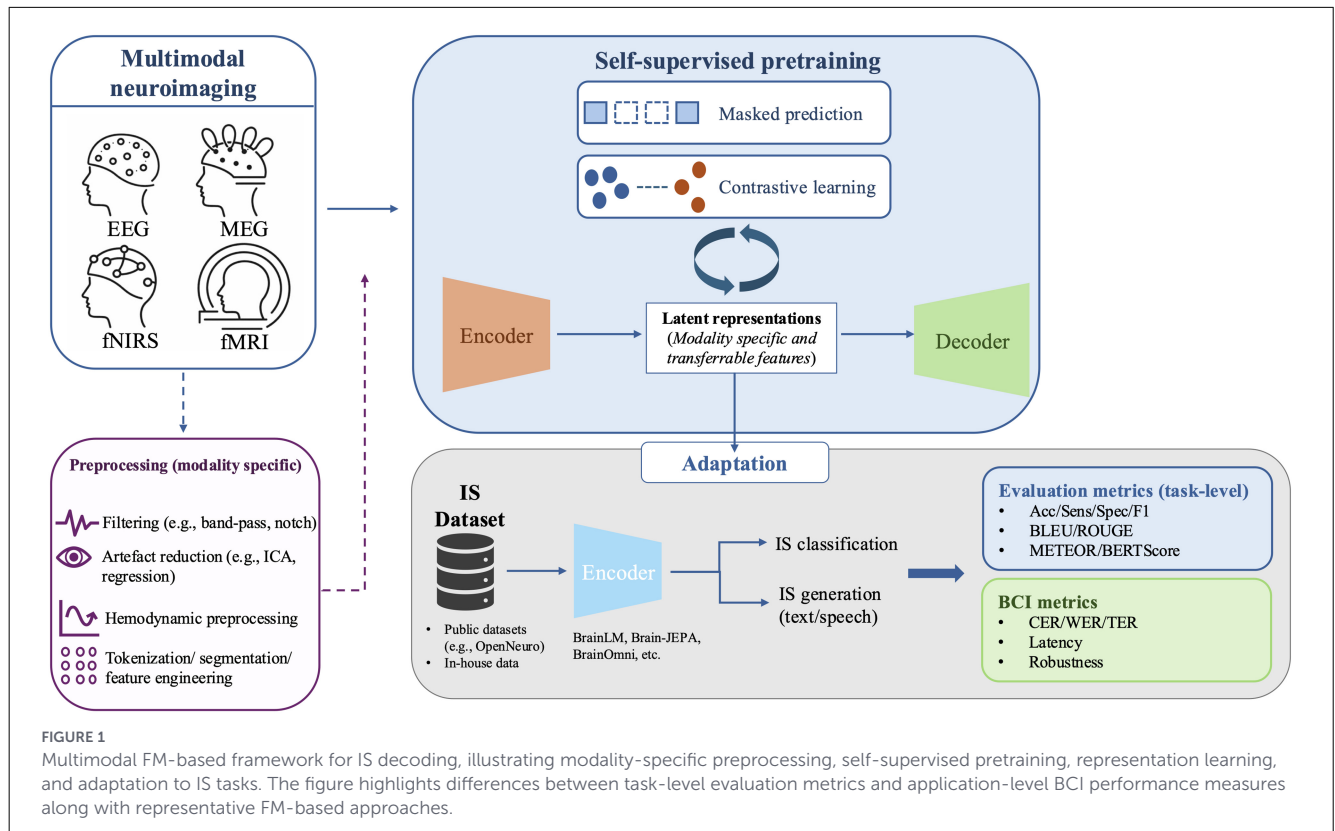
This mini review summarizes recent representative studies employing FM-based approaches using non-invasive neuroimaging modalities and their potential applicability to IS decoding. Relevant studies were identified through major databases (e.g., IEEE Xplore, PubMed, Scopus, and Google Scholar), focusing on publications from 2022 to 2025. As self-supervised learning (SSL) underpins most FMs, SSL-based non-FM approaches were also included. We highlight key developments, challenges, and future directions for advancing robust and transferable IS decoding systems.

2 Framework of foundation models for neural signal processing

FMs have emerged as a major paradigm in artificial intelligence (AI), enabling learning from large-scale unlabeled data and adaptation to multiple downstream tasks (Bommasani et al., 2021). [Figure 1](#) illustrates the overall framework of FMs for neural signal processing, including preprocessing (if applied), self-supervised pretraining, downstream adaptation, and task-dependent evaluation metrics.

- **Input data:** The models are trained on broad and often unstructured datasets to capture a general representation. More specifically, brain-signal datasets are high-dimensional, temporally rich, exhibit high noise, and vary across subjects and data acquisition conditions (Zhou et al., 2025b). Although FMs can reduce reliance on expensive manual annotation, neural recordings (e.g., EEG and fMRI) are highly irregular and dynamic, making it difficult to impose a consistent structure for pretraining (Craik et al., 2019; Wang et al., 2025a).
- **Self-supervised learning:** FMs build upon a training strategy called SSL, which leverages intrinsic information in the data by creating pretext tasks, either generative or contrastive, such as predicting masked inputs or reconstructing corrupted signals to make the model learn informative latent representations. FMs also adopt another subclass of SSL training strategy, contrastive learning, which aims to learn discriminative representations of data by using negative and positive pairs (Chen et al., 2020). FMs are trained on broad datasets (large language models (LLMs), videos, images, brain signals, structured data) with SSL or unsupervised techniques to be adaptable to downstream tasks (Bommasani et al., 2021).
- **Architecture:** Designing network architecture is another key factor that influences FM's ability to capture and encode relevant information from the raw signal (Bommasani et al., 2021). The neural-signal FMs aim to (i) capture local signal patterns, (ii) model cross-channel interactions while remaining robust to noise and artifacts, and (iii) scale to temporal contexts to learn broader dependencies.

Neural-signal FMs usually employ tokenization layers, local representation learning block, spatiotemporal attention mechanisms, and downstream task heads (Kuruppu et al., 2025). To handle the non-stationary nature of neural signals, FMs can adopt a signal segmentation method to transform continuous time series into structured representations, a process referred to as tokenization. Depending on model architecture and signal, tokenization strategies (window length, overlap vs. non-overlap patches, etc.) can be optimized for increasing computational efficiency and better modeling performance (Gu et al., 2025). On the other hand, tokenization provides a unified representation of raw signals and placed into a common input format, which can obscure or weaken the spatiotemporal information. To mitigate this, positional embeddings are utilized and added to the projected input to preserve temporal and spatial structure (Zhou et al., 2025b).



Neural-signal FMs are typically built using several non-linear transformer blocks that can capture both local patterns and long-range dependencies. Moreover, transformer models provide more effective model parallelism, which makes them well-suited to large-scale datasets. However, non-transformer models such as CNN or CNN-transformer hybrids can be good choices for avoiding potential overfitting through exploiting the spatial inductive bias of convolutions (Wu et al., 2021). Additionally, neural-signal FMs may adopt different configurations: encoder-only models are conceptually described as comprising two high-level functional components: a backbone encoder and a final fully connected layer. The backbone acts as a feature extractor, transforming the inputs into lower-dimension representations. These embeddings are then passed to a task head, typically a single fully connected layer, which produces task-specific outputs such as classification or regression (Kuruppu et al., 2025). In contrast, encoder-decoder hybrid models include an additional component: a decoder module that generates outputs conditioned on encoder representations, usually well-suited for generative or reconstruction-based tasks (e.g., neural signal-to-text or neural signals-to-speech embeddings) (He et al., 2022; Lewis et al., 2020).

- **Model adaptation:** The pretrained backbone enables reuse and transfer of neural-signal representations across multiple downstream tasks with related semantic structure. The degree of adaptation can be adjusted depending on data availability and domain shift; for instance, the backbone may be kept

fully frozen and used only as a feature extractor or updated using parameter-efficient fine-tuning (PEFT) approaches (Houlsby et al., 2019) that update only a small subset of parameters (Kottapalli et al., 2025).

3 Foundation models in non-invasive neuroimaging and inner speech decoding

Recent works have increasingly applied FM principles to non-invasive neuroimaging modalities, including EEG, MEG, fMRI, and fNIRS. In this review, we adopt an operational definition of FMs based on large-scale pretraining, SSL paradigm, and transferability across downstream tasks, categorizing models as FMs or SSL/non-FM (Table 1). While only a subset directly targets IS decoding, others were included for their transferable representations. Additionally, IS studies vary substantially in task complexity, ranging from closed-set classification to continuous open-vocabulary generation, which is substantially more challenging for non-invasive IS decoding.

In electrophysiological modalities such as EEG and MEG, large-scale pre-trained models including Large Brain Model (LaBraM) (Jiang et al., 2024) and Large Brain Language Model (LBLM) (Zhou et al., 2025a) have aimed to learn generic and transferable representations from heterogeneous multi-subject datasets. These models are adapted to tasks such as classification and IS decoding (Zhou et al., 2025a). Building on this direction, transformer-based

TABLE 1 Summary of representative IS decoding and SSL/FM-based approaches across neuroimaging modalities.

Study	Modality	Task type	Task/application	Performance	Notes
SSL/non-FM					
Shuzo et al. (2025)	EEG	Classification	IS decoding (Tasin et al., 2024)	Acc = 16.5%; F1 = 0.137	Transformer (BERT based); closed-set (6-class IS); cross-subject; low performance
Wang et al. (2024)	EEG	Generation	EEG-to-text decoding (Hollenstein et al., 2018, 2023)	BLEU-1: 42.09%; BLEU-4: 8.99%; ROUGE-1 F1: 32.61%	Multi-stream transformer; open vocabulary (reading-based EEG-to-text, not specific to IS)
Zhou et al. (2025a)	EEG	Representation learning and classification	IS decoding	Semantic acc: 47.0%, word acc: 39.6%	FM-like characteristics; FSTP pre-training; closed-set vocabulary
Lee et al. (2023)	EEG	Generation	IS decoding	CER = 68.26%; MOS = 2.78; RMSE = 0.175	Spoken EEG supervision (domain adaptation); supports unseen word generation (limited performance); closed-set (12-class IS)
Jayalath et al. (2024)	MEG	Decoding	Perceived speech (Armeni et al., 2022; Gwilliams et al., 2023)	ROC AUC = 0.705	SSL-based transformer; perceived speech classification (closed-set); near surgical-level performance
Millet et al. (2022)	fMRI	Representation learning and decoding	Speech perception (Nastase et al., 2019; Li et al., 2021)	Correlation (R) ~0.20	SSL (wav2vec 2.0); not IS-specific; listening task
FM					
Jiang et al. (2024)	EEG	Representation learning and classification	Abnormality detection and event classification (Obeid and Picone, 2016)	Task-dependent; improves downstream performance	Large scale pre-trained EEG FM
Caro et al. (2024)	fMRI	Representation learning and classification	Clinical variable and brain state prediction (Miller et al., 2016; Elam et al., 2021)	Task-dependent; improves performance	Not IS-specific
Dong et al. (2024)	fMRI	Multitask and representation learning	Clinical variable and neurodegenerative disease prediction (Miller et al., 2016; Jack Jr et al., 2008)	Task-dependent; improves performance	Not IS specific; transferable
Wang et al. (2025b)	fMRI	Multitask and representation learning	Clinical classification (HD-200 Consortium, 2012; Jack Jr et al., 2008; Marek et al., 2011; Di Martino et al., 2014) and Asian participants (MACC)	Task-dependent; improves performance	Not IS-specific
Wang et al. (2025a)	fMRI	Multitask and representation learning	Phenotype and diagnosis prediction (Van Essen et al., 2013; Meling et al., 2024)	Task-dependent; improves performance	Not IS-specific; transferrable
Wei et al. (2025)	fMRI	Representation learning and generation	Phenotype and fMRI-to-text (Elam et al., 2021; Miller et al., 2016; Karcher and Barch, 2021)	Task-dependent; strong generalization	Not IS specific; language-aligned; potential relevance to IS
Jung et al. (2025)	fNIRS	Representation learning and classification	Cognitive classification	Task-dependent; acc ~0.67–0.79	Data-efficient; not IS specific; transferrable

(Continued)

TABLE 1 (Continued)

Study	Modality	Task type	Task/application	Performance	Notes
Zhang S. et al. (2024)	fNIRS	Generation	IS decoding (4 participants)	BLEU-1 = 0.25; BERTScore = 0.88; METEOR = 0.17; ROUGE-L = 0.21; WER = 0.84 (mean across subjects)	Small sample size; prompt-tuned LLM integration; continuous and open-vocabulary; limited lexical acc and moderate semantic decoding
Ferrante et al. (2025)	EEG, MEG, and fMRI	Representation learning and decoding	Decoding, encoding, modality conversion	Task-dependent; improves cross-modal performance	Not IS specific; transferable across modalities
Xiao et al. (2025)	EEG and MEG	Representation learning and classification	Multiple datasets (clinical, motor, multimodal)	Balanced acc; outperforms baselines	Not IS specific; large-scale multimodal FM

The table distinguishes between SSL/non-FM and FM approaches, highlighting differences in task types, applications, datasets, and performance metrics.

Acc, Accuracy; AUC-PR, area under the precision–recall curve; ROC AUC, receiver operating characteristic area under the curve; RMSE, root mean squared error; CER, character error rate; MOS, mean opinion score; MACC, Memory, Ageing and Cognition Centre; LLM, large language model; WER, word error rate; FSTP, future spectro-temporal prediction; BLEU, Bilingual Evaluation Understudy; BERTScore, Bidirectional Encoder Representations from Transformers Score; METEOR, Metric for Evaluation of Translation with Explicit Ordering; ROUGE, Recall-Oriented Understudy for Gisting Evaluation.

architectures have also been explored for EEG-based IS decoding. For example, Shuzo et al. (2025) applied a pre-trained lightweight BERT model (Devlin et al., 2019) for six-class IS classification; however, the reported performance (accuracy = 0.165, F1 = 0.137) is close to chance level, indicating limited performance in EEG-based IS decoding. Other approaches, such as contrastive EEG–text masked autoencoders (CET-MAE) (Wang et al., 2024), have introduced language-aligned representation learning, using EEG embeddings to map into shared semantic spaces with textual representations. Recent pre-training paradigms, including Future Spectro-Temporal Prediction (Zhou et al., 2025a), and generative frameworks such as NeuroTalk (Lee et al., 2023), have further explored potentially transferable representation learning for IS, primarily in limited-vocabulary settings. A complementary MEG study has also suggested that, rather than using handcrafted features and small supervised models, scaling data and SSL may improve cross-subject speech decoding performance (Jayalath et al., 2024). Although the method was demonstrated primarily for decoding perceived speech rather than IS, the study can contribute to the critical step of cross-subject decoding of IS technologies.

In fMRI research, similar trends toward representation learning have been proposed. Models such as BrainLM (Caro et al., 2024), Brain-JEPA (Dong et al., 2024), SLIM-Brain (Wang et al., 2025b), and NeuroSTORM (Wang et al., 2025a) have marked a shift from task-specific prediction toward learning general cortical representations. In this direction, fMRI-LM (Wei et al., 2025) has incorporated large-scale, multi-subject pretraining aligned with pretrained language models, aiming to improve transferability across subjects. Converging evidence from speech models, Millet et al. demonstrated that SSL speech models, such as wav2vec 2.0 (Baevski et al., 2020), exhibited hierarchical representations that align with distributed cortical speech processing patterns (Millet et al., 2022). These findings suggested that the proposed models capture biologically meaningful and hierarchically structured speech representations.

Emerging FM approaches have been explored in hemodynamic neuroimaging beyond fMRI. For example, fNIRS-based models (Jung et al., 2025) have demonstrated that scalable pre-training strategies can extend to hemodynamic modalities. Recent systems, such as MindSpeech (Zhang S. et al., 2024), have employed pretrained language models to enable continuous, open-vocabulary, and semantically informed decoding. However, the reported performance indicates low lexical accuracy despite moderate semantic similarity, suggesting that current performance remains below clinically viable thresholds for reliable communication. These limitations highlight the constraints of single-modality decoding and motivate a shift toward unified frameworks that integrate complementary non-invasive neural modalities (Ferrante et al., 2025). For instance, BrainOmni has been designed to generalize across EEG and MEG recordings, learning shared spatio-temporal representations (Xiao et al., 2025). The model has aimed to capture modality-invariant neural dynamics across EEG and MEG. This cross-modality framework may enable knowledge transfer for IS decoding.

4 Discussion

Despite rapid advances in FM approaches for BCI, several structural, methodological, and practical obstacles persist, limiting scalable and reliable IS decoding.

4.1 Scale, computational constraints, and data heterogeneity

FMs benefit from massive, diverse datasets to learn latent structural patterns in neural data through self-supervised objectives. However, in neuroimaging research, data collection

remains limited and costly (Wang et al., 2025c). In addition, training high-capacity spatiotemporal models on neuroimaging data is computationally demanding.

Beyond scale, neural data are inherently heterogeneous and non-stationary. Variability in acquisition protocols, scanner parameters, preprocessing pipelines, and experimental paradigms complicates data aggregation and large-scale pretraining. Neural recordings are also vulnerable to artifacts, further challenging robustness. These issues become more pronounced in multimodal settings, where EEG, MEG, fMRI, and fNIRS require harmonizing signals with fundamentally different spatial and temporal properties (Zhou et al., 2025b).

Recent large-scale datasets, such as LibriBrain, comprising over 50 h of MEG recordings during speech processing (Özdoğan et al., 2025) and MOUS, which includes multimodal recordings from 204 subjects (Schoffelen et al., 2019), represent important steps toward scalable neural speech modeling. However, datasets specifically designed for IS are typically limited to single-modality recordings and relatively small cohorts (Nieto et al., 2022) with only a few studies acquiring simultaneous multimodal neural data (Cooney et al., 2022; Liwicki et al., 2025). Although recent efforts such as the Chisco dataset have introduced larger-scale EEG-based imagined speech corpora comprising semantically diverse daily expressions (Zhang Z. et al., 2024), IS datasets remain limited in subject diversity and recording modalities.

Moreover, IS datasets remain limited in language diversity (Simistira Liwicki et al., 2023; Zhang Z. et al., 2024). This limitation is particularly important, as language shapes the neural representation of IS. For instance, linguistic properties such as morphological complexity, tonal structure, and word segmentation may produce distinct cortical activation patterns (Dash et al., 2020). Efforts to extend IS decoding to multiple languages underscore the importance of linguistic diversity in datasets and the systematic analysis of cross-linguistic differences in neural signals to identify universal neural correlates of covert speech (Almufareh et al., 2025). Addressing these challenges will require standardized data sharing (e.g., OpenNeuro), harmonized preprocessing pipelines, standardized data formats such as Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2016), consistent experimental paradigms, and scalable training strategies tailored to neuroimaging data.

Overcoming these limitations is essential for enabling effective large-scale FM, facilitating cross-subject and cross-lingual generalization, and ultimately learning robust, invariant, and generalizable neural representations for IS decoding.

4.2 Neurophysiological constraints of IS and implications for FMs

Decoding the content of IS, i.e., internally generated speech from complex neural activity, presents several task-specific challenges. First, neural signals associated with IS are typically weaker and noisier than those observed in overt or attempted speech (Almufareh et al., 2025). Although IS has been

associated with internal motor predictions such as efference copies (Whitford et al., 2017), the absence of overt articulation limits the availability of external timing markers and the predictability of neural signals (Wang et al., 2025c).

Furthermore, non-invasive IS decoding is particularly constrained by low signal-to-noise ratios (e.g., in scalp EEG), the distributed and complex cortical encoding of speech, and substantial inter-subject variability. From an FM perspective, these challenges highlight the need for SSL approaches capable of learning invariant spatiotemporal representations from weak, noisy, and non-stationary neural signals. In addition, developing robust decoders that can disentangle variability in neural representations arising from neurological conditions (e.g., stroke and LIS), as well as differences in cognitive profiles, remains a central challenge in the field.

4.3 Evaluation metrics and interpretability

Traditional performance metrics, such as classification accuracy and F1 score, are suitable for closed-set classification tasks but are not sufficient to capture semantic fidelity in sentence-level decoding. In generative frameworks, metrics such as Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2001) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) provide assessments of exact textual overlap. However, these metrics do not fully capture semantic fidelity. For more semantically informed evaluation, the Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee and Lavie, 2005) and BERTscores (Zhang et al., 2019) have been employed. Furthermore, practical BCI systems are typically assessed by performance measures such as bit rate, character error rate (CER), word error rate (WER), average token error rate (TER), and latency, which reflect communication efficiency and real-time usability (He et al., 2026). Nevertheless, the lack of standardized evaluation protocols that reliably reflect deeper conceptual meaning and provide a concise summary of model performance hinders fair comparisons in IS decoding research.

Another critical bottleneck lies in the interpretability of FM-based decoders. Deep learning models capture complex non-linear patterns that are difficult to interpret in neural terms and therefore often work as “black-box” systems. Combined with cortical complexity and inter-subject variability, this limits explainability and raises concerns about trust and ethical deployment (Doshi-Velez and Kim, 2017). Methods such as attention and saliency visualizations offer some degree of insight about models’ decisions; however, they often do not provide enough detail to understand the mechanism behind a model’s decisions and lack of neuroscientific ground (Rudin, 2019). More neuroscientific approaches, such as Representational Similarity Analysis (RSA), can compare the similarity structure of the model embeddings with cortical activity patterns (Kriegeskorte, 2008). Developing inherently interpretable frameworks that provide transparent outputs and clear feedback while aligning with model-derived neural representation with neuroscience principles remains an important challenge.

4.4 Continuous and real-time IS decoding

Real-world BCI applications require FM-based decoders capable of processing weak, non-stationary neural signals in real time. However, the absence of overt behavioral anchors in IS complicates onset detection and poses a fundamental challenge for continuous, open-vocabulary decoding (Martin et al., 2014). Moreover, non-invasive modalities provide indirect and spatially coarse measurements of neural activity, limiting their ability to capture high-frequency articulatory representations associated with speech production. Designing FMs that balance computational efficiency, robustness, and adaptability, therefore, remains an open challenge for scalable BCI systems.

In contrast, invasive BCI systems based on ECoG have demonstrated continuous speech decoding (Anumanchipalli et al., 2019) and, in some cases, near real-time operation (Metzger et al., 2023; Moses et al., 2021), achieving substantially higher performance and larger vocabularies. These systems leverage high-frequency cortical activity directly linked to speech production, resulting in significantly higher signal fidelity. Although most results have been derived from overt or attempted speech rather than IS, they have established a practical upper bound for decoding performance. This highlights a gap between invasive and non-invasive approaches, reflecting differences in signal accessibility and posing a key limitation for achieving continuous IS decoding with non-invasive FMs.

4.5 Ethical and privacy considerations

Ethical and privacy considerations are central to the development of IS decoding technologies. Neural data are highly sensitive and subject to strict privacy and institutional constraints. Unlike many other biomedical signals, brain-derived data may reveal not only speech content but also aspects of a participant's mental state (Ienca and Andorno, 2017). This concern is particularly important for IS decoding, as IS is widely described as an intrinsically private mental process through which individuals plan, reflect, and encode representations that guide behavior (Alderson-Day and Fernyhough, 2015). Accordingly, strong privacy safeguards and transparent frameworks are essential. Emerging technical solutions, such as federated learning, can enable collaborative model training without centralized data sharing, while biologically informed synthetic data augmentation may reduce reliance on raw neural recordings (Kwon and Shin, 2026). Transparency and explainability are also critical for strengthening public trust in IS technologies and addressing concerns related to “mind reading.” Robust frameworks for informed consent, user agency, and clearly defined privacy boundaries remain essential prerequisites for responsible deployment (Wang et al., 2025c).

As these technologies transition to real-world applications, broader ethical and societal considerations become increasingly important. Issues of access, regulation, and responsible use must be carefully addressed, particularly regarding potential biases, unequal access, and the amplification of existing

social inequalities. Addressing these challenges will require interdisciplinary collaboration among engineers, ethicists, clinicians, and policymakers to ensure that technological advances are aligned with appropriate governance frameworks and responsible deployment (Almufareh et al., 2025).

In conclusion, FMs are increasingly applied to non-invasive neuroimaging for IS decoding. Through large-scale self-supervised pretraining and transferable representations, they offer a promising alternative to data-intensive supervised approaches. These models may improve cross-subject generalization, capture contextual dependencies, and mitigate data scarcity. However, key challenges remain, including limited dataset scale, neurophysiological constraints, inconsistent evaluation metrics (especially for semantic decoding), interpretability, and ethical concerns. Achieving continuous, real-time IS decoding with robust, transparent, and privacy-preserving systems remains difficult. Bridging advances in FMs, neuroscience, and human-computer interaction will be essential for developing scalable and trustworthy IS decoding systems.

Author contributions

ES-A: Conceptualization, Writing – original draft, Writing – review & editing. RS: Writing – original draft, Writing – review & editing. DC: Writing – original draft, Writing – review & editing. SV: Writing – original draft, Writing – review & editing. FS: Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the Kempe Foundations under project number JCSMK25-0068. The publication fee is covered by Luleå University of Technology.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. Generative AI was used in the preparation of this manuscript for language editing and clarity improvement. The author(s) take full responsibility for the content of the manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of

artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the

reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnhum.2026.1838064/full#supplementary-material>

References

- Alderson-Day, B., and Fernyhough, C. (2015). Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychol. Bull.* 141, 931–965. doi: 10.1037/bul0000021
- Almufareh, M. F., Kausar, S., Humayun, M., Tehsin, S., and Farooq, A. (2025). Inner speech decoding: a comprehensive review. *Wiley Interdiscipl. Rev.: Cogn. Sci.* 16:e70016. doi: 10.1002/wcs.70016
- Anumanchipalli, G. K., Chartier, J., and Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature* 568, 493–498. doi: 10.1038/s41586-019-1119-1
- Armeni, K., Güçlü, U., van Gerven, M., and Schoffelen, J.-M. (2022). A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Sci. Data* 9:278. doi: 10.1038/s41597-022-01382-7
- Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). Wav2vec 2.0: a framework for self-supervised learning of speech representations. Version 3. *arXiv [preprint]*. doi: 10.48550/arXiv.2006.11477
- Banerjee, S., and Lavie, A. (2005). "Meteor: an automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (Kerrville, TX: Association for Computational Linguistics), 65–72.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. Version 3. *arXiv [preprint]* *arXiv:2108.07258*. doi: 10.48550/arXiv.2108.07258
- Caro, J. O., Fonseca, A., Averill, C., Rizvi, S., Rosati, M., Averill, C., et al. (2024). "A foundation model for brain activity recordings," in *ICLR 2024 Conference* (OpenReview.net). doi: 10.1101/2023.09.12.557460
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning* (Cambridge, MA: PMLR), 1597–1607.
- Chengaiyan, S., Retnapandian, A. S., and Anandan, K. (2020). Identification of vowels in consonant-vowel-consonant words from speech imagery based eeg signals. *Cogn. Neurodyn.* 14, 1–19. doi: 10.1007/s11571-019-09558-5
- Cooney, C., Folli, R., and Coyle, D. (2022). A bimodal deep learning architecture for EEG-fNIRS decoding of overt and imagined speech. *IEEE Trans. Biomed. Eng.* 69, 1983–1994. doi: 10.1109/TBME.2021.3132861
- Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *J. Neural Eng.* 16:031001. doi: 10.1088/1741-2552/ab0ab5
- Dash, D., Ferrari, P., and Wang, J. (2020). Decoding imagined and spoken phrases from non-invasive neural (meg) signals. *Front. Neurosci.* 14:290. doi: 10.3389/fnins.2020.00290
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers)* (Kerrville, TX: Association for Computational Linguistics), 4171–4186. doi: 10.18653/v1/N19-1423
- Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78
- Dong, Z., Li, R., Wu, Y., Nguyen, T. T., Chong, J. S., Ji, F., et al. (2024). Brain-JEPA: brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Adv. Neural Inf. Process. Syst.* 37, 86048–86073. doi: 10.52202/079017-2732
- Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. Version 2. *arXiv [preprint]* *arXiv:1702.08608*. doi: 10.48550/arXiv.1702.08608
- Elam, J. S., Glasser, M. F., Harms, M. P., Sotiropoulos, S. N., Andersson, J. L., Burgess, G. C., et al. (2021). The human connectome project: a retrospective. *Neuroimage* 244:118543. doi: 10.1016/j.neuroimage.2021.118543
- Ferrante, M., Boccatto, T., Rashkov, G., and Toschi, N. (2025). Towards neural foundation models for vision: aligning EEG, MEG, and fMRI representations for decoding, encoding, and modality conversion. *Inform. Fus.* 126:103650. doi: 10.1016/j.inffus.2025.103650
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.44
- Gu, X., Shu, Y., Han, J., Liu, Y., Liu, Z., Anibal, J., et al. (2025). Foundation models for biosignals: a survey. *Authorea Preprints*. doi: 10.36227/techrxiv.175606236.62808131/v1
- Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., et al. (2024). A survey on self-supervised learning: algorithms, applications, and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 9052–9071. doi: 10.1109/TPAMI.2024.3415112
- Gwilliams, L., Flick, G., Marantz, A., Pylkkänen, L., Poeppel, D., and King, J.-R. (2023). Introducing MEG-MASC a high-quality magnetoencephalography dataset for evaluating natural speech processing. *Sci. Data* 10:862. doi: 10.1038/s41597-023-02752-5
- HD-200 Consortium (2012). The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Front. Syst. Neurosci.* 6:62. doi: 10.3389/fnsys.2012.00062
- He, D., Siok, W. T., and Wang, N. (2026). Toward robust, reproducible, and widely accessible intracranial language brain-computer interfaces: a comprehensive review of neural mechanisms, hardware, algorithms, evaluation, clinical pathways and future directions. Version 2. *arXiv [preprint]* *arXiv:2603.12279*. doi: 10.48550/arXiv.2603.12279
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Los Alamitos, CA: IEEE Computer Society), 16000–16009. doi: 10.1109/CVPR52688.2022.01553
- Hernández-Del-Toro, T., Reyes-García, C. A., and Villasenor-Pineda, L. (2021). Toward asynchronous EEG-based BCI: detecting imagined words segments in continuous EEG signals. *Biomed. Signal Process. Control* 65:102351. doi: 10.1016/j.bspc.2020.102351
- Hollenstein, N., Rotsztein, J., Troendle, M., Pedroni, A., Zhang, C., and Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Sci. Data* 5:180291. doi: 10.1038/sdata.2018.291
- Hollenstein, N., Tröndle, M., Plomecka, M., Kieglend, S., Özyurt, Y., Jäger, L. A., et al. (2023). The zuco benchmark on cross-subject reading task classification with eeg and eye-tracking data. *Front. Psychol.* 13:1028824. doi: 10.3389/fpsyg.2022.1028824
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., et al. (2019). "Parameter-efficient transfer learning for NLP" in *International Conference on Machine Learning* (Cambridge, MA: PMLR), 2790–2799.
- Ienca, M., and Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sci. Soc. Policy* 13:5. doi: 10.1186/s40504-017-0050-1
- Jack, C. R. Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* 27, 685–691. doi: 10.1002/jmri.21049

- Jayalath, D., Landau, G., Shillingford, B., Woolrich, M., and Jones, O. P. (2024). The brain's bitter lesson: scaling speech decoding with self-supervised learning. Version 5. *arXiv [preprint] arXiv:2406.04328*. doi: 10.48550/arXiv.2406.04328
- Jiang, W.-B., Zhao, L.-M., and Lu, B.-L. (2024). Large brain model for learning generic representations with tremendous EEG data in BCI. Version 1. *arXiv [preprint] arXiv:2405.18765*. doi: 10.48550/arXiv.2405.18765
- Jung, E., Lee, H., and An, J. (2025). "fNIRS foundation model for few-shot based fNIRS classification," in *2025 13th International Conference on Brain-Computer Interface (BCI)* (Piscataway, NJ: IEEE), 1–4. doi: 10.1109/BCI65088.2025.10931275
- Karcher, N. R., and Barch, D. M. (2021). The ABCD study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology* 46, 131–142. doi: 10.1038/s41386-020-0736-6
- Komeiji, S., Mitsuhashi, T., Iimura, Y., Suzuki, H., Sugano, H., Shinoda, K., et al. (2024). Feasibility of decoding covert speech in ECoG with a Transformer trained on overt speech. *Sci. Rep.* 14:11491. doi: 10.1038/s41598-024-62230-9
- Kottapalli, S. R. K., Hubli, K., Chandrashekhar, S., Jain, G., Hubli, S., Botla, G., et al. (2025). Foundation models for time series: a survey. Version 1. *arXiv [preprint] arXiv:2504.04011*. doi: 10.48550/arXiv.2504.04011
- Kriegeskorte, N. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008
- Kuruppu, G., Wagh, N., Kremen, V., and Varatharajah, Y. (2025). EEG foundation models: a critical review of current progress and future directions. *J. Neural Eng.* 23:021001. doi: 10.1088/1741-2552/ae4455
- Kwon, J., and Shin, Y. (2026). Foundation models for neural signal decoding: EEG-centered perspectives toward unified representations. *Eur. J. Neurosci.* 63:e70376. doi: 10.1111/ejn.70376
- Lee, Y.-E., Lee, S.-H., Kim, S.-H., and Lee, S.-W. (2023). "Towards voice reconstruction from EEG during imagined speech," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37 (Washington, DC: AAAI Press), 6030–6038. doi: 10.1609/aaai.v37i5.25745
- Lesaja, S., Stuart, M., Shih, J. J., Soroush, P. Z., Schultz, T., Manic, M., et al. (2022). Self-supervised learning of neural speech representations from unlabeled intracranial signals. *IEEE Access* 10, 133526–133538. doi: 10.1109/ACCESS.2022.3230688
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2020). "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Kerrville, TX: Association for Computational Linguistics), 7871–7880. doi: 10.18653/v1/2020.acl-main.703
- Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R. N., et al. (2021). Le Petit Prince: a multilingual fMRI corpus using ecological stimuli. *bioRxiv [preprint]*. doi: 10.1101/2021.10.02.462875
- Lin, C.-Y. (2004). "ROUGE: a package for automatic evaluation of summaries," in *Text Summarization Branches Out* (Barcelona: Association for Computational Linguistics), 74–81.
- Liwicki, F. S., Saini, R., Chakladar, D. D., Rakesh, S., Gupta, V., Eriksson, J., et al. (2025). Simultaneous electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) data during an inner speech task. *Data Brief* 63:112258. doi: 10.1016/j.dib.2025.112258
- Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., et al. (2011). The parkinson progression marker initiative (PPMI). *Progress Neurobiol.* 95, 629–635. doi: 10.1016/j.neurobio.2011.09.005
- Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., et al. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Front. Neuroeng.* 7:14. doi: 10.3389/fneng.2014.00014
- Martin, S., Brunner, P., Iturrate, I., Millán, J., d., R., Schalk, G., et al. (2016). Word pair classification during imagined speech using direct brain recordings. *Sci. Rep.* 6:25803. doi: 10.1038/srep25803
- Martin, S., Iturrate, I., Millán, J. R., Knight, R. T., and Pasley, B. N. (2018). Decoding inner speech using electrocorticography: progress and challenges toward a speech prosthesis. *Front. Neurosci.* 12:422. doi: 10.3389/fnins.2018.00422
- Meling, D., Egger, K., Aicher, H. D., Jareño Redondo, J., Mueller, J., Dornbierer, J., et al. (2024). Meditating on psychedelics. A randomized placebo-controlled study of DMT and harmine in a mindfulness retreat. *J. Psychopharmacol.* 38, 897–910. doi: 10.1177/02698811241282637
- Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., et al. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* 620, 1037–1046. doi: 10.1038/s41586-023-06443-4
- Miller, K. L., Alfaro-Almagro, F., Bangarter, N. K., Thomas, D. L., Yacoub, E., Xu, J., et al. (2016). Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nat. Neurosci.* 19, 1523–1536. doi: 10.1038/nn.4393
- Millet, J., Caucheteux, C., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., et al. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *Adv. Neural Inf. Process. Syst.* 35, 33428–33443. doi: 10.52202/068431-2422
- Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., et al. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *N. Engl. J. Med.* 385, 217–227. doi: 10.1056/NEJMoa2027540
- Nastase, S., Liu, Y.-F., Hillman, H., Zadbod, A., Hasenfratz, L., Keshavarzian, N., et al. (2019). Narratives: fMRI data for evaluating models of naturalistic language comprehension. *bioRxiv [preprint]*. doi: 10.1101/2020.12.23.424091
- Nieto, N., Peterson, V., Rufiner, H. L., Kamienskowski, J. E., and Spies, R. (2022). Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Sci. Data* 9:52. doi: 10.1038/s41597-022-01147-2
- Obeid, I., and Picone, J. (2016). The temple university hospital EEG data corpus. *Front. Neurosci.* 10:196. doi: 10.3389/fnins.2016.00196
- Özdoğan, M., Landau, G., Elvers, G., Jayalath, D., Somaiya, P., Mantegna, F., et al. (2025). LibriBrain: over 50 hours of within-subject MEG to improve speech decoding methods at scale. Version 1. *arXiv [preprint] arXiv:2506.02098*. doi: 10.48550/arXiv.2506.02098
- Pan, H., Li, Z., Tian, C., Wang, L., Fu, Y., Qin, X., et al. (2023). The LightGBM-based classification algorithm for Chinese characters speech imagery BCI system. *Cogn. Neurodyn.* 17, 373–384. doi: 10.1007/s11571-022-09819-w
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics - ACL '02* (Philadelphia, PA: Association for Computational Linguistics), 311. doi: 10.3115/1073083.1073135
- Pei, X., Barbour, D. L., Leuthardt, E. C., and Schalk, G. (2011). Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *J. Neural Eng.* 8:046028. doi: 10.1088/1741-2560/8/4/046028
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x
- Saha, P., and Fels, S. (2019). "Hierarchical deep feature learning for decoding imagined speech from EEG," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33 (Washington, DC: AAAI Press), 10019–10020. doi: 10.1609/aaai.v33i01.330110019
- Schoffelen, J.-M., Oostenveld, R., Lam, N. H., Uddén, J., Hultén, A., and Hagoort, P. (2019). A 204-subject multimodal neuroimaging dataset to study language processing. *Sci. Data* 6:17. doi: 10.1038/s41597-019-0020-y
- Schultz, T., Wand, M., Hueber, T., Krusienski, D. J., Herff, C., and Brumberg, J. S. (2017). Biosignal-based spoken communication: a survey. *IEEE/ACM Trans. Audio Speech Lang. Process.* 25, 2257–2271. doi: 10.1109/TASLP.2017.2752365
- Shah, U., Alzubaidi, M., Mohsen, F., Abd-Alrazaq, A., Alam, T., and Househ, M. (2022). The role of artificial intelligence in decoding speech from EEG signals: a scoping review. *Sensors* 22:6975. doi: 10.3390/s22186975
- Shergill, S. S., Bullmore, E., Brammer, M., Williams, S., Murray, R., and McGuire, P. (2001). A functional study of auditory verbal imagery. *Psychol. Med.* 31, 241–253. doi: 10.1017/S003329170100335X
- Shuzo, M., Hiramoto, R., Ishigaki, R., Ando, S., and Sakai, M. (2025). Development of an EEG-based silent speech recognition model on the native Arabic silent speech dataset using light BERT architecture. *Int. J. Act. Behav. Comput.* 2025, 1–16. doi: 10.60401/ijabc.116
- Simistira Liwicki, F., Gupta, V., Saini, R., De, K., Abid, N., Rakesh, S., et al. (2023). Bimodal electroencephalography-functional magnetic resonance imaging dataset for inner-speech recognition. *Sci. Data* 10:378. doi: 10.1038/s41597-023-02286-w
- Simistira Liwicki, F., Gupta, V., Saini, R., De, K., and Liwicki, M. (2022). Rethinking the methods and algorithms for inner speech decoding and making them reproducible. *NeuroSci* 3, 226–244. doi: 10.3390/neurosci3020017
- Tasin, S. M., Chowdhury, M. E. H., Pedersen, S., Chabbouh, M., Bushnaq, D., Aljindi, R., et al. (2024). Ensemble machine learning model for inner speech recognition: a subject-specific investigation. Version Number: 1. *arXiv [preprint] arXiv:2412.17824*. doi: 10.48550/arXiv.2412.17824
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041
- Vygotsky, L. S. (1987). Thinking and speech. *Collected Works LS Vygotsky* 1, 39–285.
- Wang, C., Jiang, Y., Peng, Z., Li, C., Bang, C., Zhao, L., et al. (2025a). Towards a general-purpose foundation model for fMRI analysis. Version 2. *arXiv [preprint] arXiv:2506.11167*. doi: 10.48550/arXiv.2506.11167
- Wang, J., Song, Z., Ma, Z., Qiu, X., Zhang, M., and Zhang, Z. (2024). "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)* (Kerrville, TX: Association for Computational Linguistics), 7278–7292. doi: 10.18653/v1/2024.acl-long.393
- Wang, M., Xia, J., Ye, W., Liu, E., Peng, K., Feng, J., et al. (2025b). SLIM-brain: a data- and training-efficient foundation model for fMRI data analysis. Version 3. *arXiv [preprint] arXiv:2512.21881*. doi: 10.48550/arXiv.2512.21881

- Wang, Y., Wang, S., Cai, W., Du, C., Fan, C., Li, D., et al. (2025c). Representation, alignment, and generation: a comprehensive survey of foundation models for non-invasive brain decoding. *bioRxiv [preprint]*. doi: 10.64898/2025.11.30.691403
- Wei, Y., Zhang, Y., Xiao, X., Qian, C., Wang, T., and Calhoun, V. D. (2025). fMRI-LM: towards a universal foundation model for language-aligned fMRI understanding. Version 4. *arXiv [preprint] arXiv:2511.21760*. doi: 10.48550/arXiv.2511.21760
- Wellington, S., Wilson, H., Liwicki, F. S., Gupta, V., Saini, R., De, K., et al. (2024). "Improving inner speech decoding by hybridisation of bimodal EEG and fMRI data," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Piscataway, NJ: IEEE), 1–5. doi: 10.1109/EMBC53108.2024.10781692
- Whitford, T. J., Jack, B. N., Pearson, D., Griffiths, O., Luque, D., Harris, A. W., et al. (2017). Neurophysiological evidence of efference copies to inner speech. *Elife* 6:e28197. doi: 10.7554/eLife.28197
- Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., et al. (2021). "CvT: introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Washington, DC: IEEE Computer Society), 22–31. doi: 10.1109/ICCV48922.2021.00009
- Xiao, Q., Cui, Z., Zhang, C., Chen, S., Wu, W., Thwaites, A., et al. (2025). BrainOmni: a brain foundation model for unified EEG and MEG signals. Version 3. *arXiv [preprint] arXiv:2505.18185*. doi: 10.48550/arXiv.2505.18185
- Zhang, S., Alam, E., Baber, J., Bianco, F., Turner, E., Chamanzar, M., et al. (2024). MindSpeech: continuous imagined speech decoding using high-density fNIRS and prompt tuning for advanced human-AI interaction. Version 1. *arXiv [preprint] arXiv:2408.05362*. doi: 10.48550/arXiv.2408.05362
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). BERTScore: evaluating text generation with BERT. Version 3. *arXiv [preprint] arXiv:1904.09675*. doi: 10.48550/arXiv.1904.09675
- Zhang, Z., Ding, X., Bao, Y., Zhao, Y., Liang, X., Qin, B., et al. (2024). Chisco: an EEG-based BCI dataset for decoding of imagined speech. *Scientific Data* 11:1265. doi: 10.1038/s41597-024-04114-1
- Zhou, J., Cao, Z., Duan, Y., Barkley, C., Leong, D., Jiang, X., et al. (2025a). "Pretraining large brain language model for active BCI: silent speech," in *Proceedings of the 33rd ACM International Conference on Multimedia* (New York, NY: Association for Computing Machinery), 5883–5892. doi: 10.1145/3746027.3754810
- Zhou, X., Liu, C., Chen, Z., Wang, K., Ding, Y., Jia, Z., et al. (2025b). Brain foundation models: a survey on advancements in neural signal processing and brain discovery. *IEEE Signal Process. Mag.* 42, 22–35. doi: 10.1109/MSP.2025.3592356