Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Large Language Models Integrated into Brain-Computer Interfaces for Communication and Control: A Systematic Review

Alon Gorenshtein

agorensh@bidmc.harvard.edu

Department of Neurology, Harvard Medical School, Boston, MA https://orcid.org/0009-0000-7542-8608

Mahmud Omar

The Windreich Department of Artificial Intelligence and Human Health, Mount Sinai Medical Center, NY, USA. https://orcid.org/0009-0001-0438-0827

Yiftach Barash

Division of Vascular and Interventional Radiology, Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical

Girish N Nadkarni

The Windreich Department of Artificial Intelligence and Human Health, Mount Sinai Medical Center, NY, USA. https://orcid.org/0000-0001-6319-4314

Eyal Klang

The Windreich Department of Artificial Intelligence and Human Health, Mount Sinai Medical Center, NY, USA. https://orcid.org/0000-0002-4567-3108

Systematic Review

Keywords: Neuro-Symbolic, AI agent, Large language model, systematic review, AI, Brain-Computer Interface, Large Language Models, Generative AI, Systematic Review, Neural Prosthetics, Neurotechnology

Posted Date: March 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8981776/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Large Language Models Integrated into Brain-Computer Interfaces for Communication and Control: A Systematic Review

Alon Gorenshtein,[1,2,3,4] Mahmud Omar,[1,2] Yiftach Barash,[5] Girish N Nadkarni,[1,2] Eyal Klang[1,2]

[1] The Windreich Department of Artificial Intelligence and Human Health, Mount Sinai Medical Center, NY, USA.
[2] The Hasso Plattner Institute for Digital Health at Mount Sinai, Mount Sinai Health System, NY, USA.
[3] Department of Neurology, Harvard Medical School, Boston, MA
[4] Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA
[5] Division of Vascular and Interventional Radiology, Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

Author Contributions: Conceptualization, AG, MO, EK, YB, GN; Methodology, AG, MO, EK; Formal Analysis, AG, MO; Data Curation, AG, MO; Writing-Original Draft Preparation, AG,MO; Writing- Review & Editing, AG, MO, EK,YB, GN; Supervision: EK
Funding: This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. Research reported in this publication was also supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880 and S10OD030463. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Competing interests: The authors declare that they have no competing interests.
Word count: 3584.

Address correspondence to:
Alon Gorenshtein MD
Beth Israel Deaconess Medical Center, Harvard Medical school
330 Brookline Ave, Boston, MA 02115
Agorensh@bidmc.harvard.edu

***Abstract*—The integration of Large Language Models (LLMs) into Brain-Computer Interfaces (BCIs) represents a paradigm shift in neural prosthetics, offering unprecedented opportunities to enhance communication rates, accuracy, and user experience. This systematic review synthesizes recent evidence on LLM-assisted BCIs for communication and control. Following PRISMA guidelines, we extracted data from nine included studies covering diverse BCI modalities (P300, SSVEP, cVEP, passive, auditory) and LLM integration paradigms, including autocomplete, post-edit correction, intent expansion, dynamic UI generation, and affective support. We detail the methodological pipelines, synthesize the quantitative outcomes (such as changes in spelling accuracy, Information Transfer Rate, and Keystroke Savings), and conduct a rigorous evaluation using established Risk of Bias tools. Results indicate that LLM integration consistently improves effective communication speed and reduces user fatigue, though the reliance on remote APIs introduces new constraints related to latency, privacy, and safety (e.g., hallucinations and unprompted corrections). We propose a taxonomy of BCI-LLM integration patterns and discuss the gap between simulated offline enhancements and real-time clinical validation. Finally, we provide standardized reporting recommendations and outline a concrete research agenda to guide the development of secure, personalized, and clinically viable LLM-assisted BCI systems.**

***Index Terms*—Neuro-Symbolic, AI agent, Large language model, systematic review, AI, Brain-Computer Interface, Large Language Models, Generative AI, Systematic Review, Neural Prosthetics, Neurotechnology.**

I. Introduction

Brain computer interfaces (BCIs) provide direct communication pathways between the brain and external devices, offering profound restitution of function for individuals with severe motor impairments.[1] Despite decades of advancement, spelling and communication BCIs continue to face significant usability hurdles, including low Information Transfer Rates (ITR), high cognitive workload, and susceptibility to decoding errors caused by electroencephalography (EEG) noise and signal non-stationarity.[2,3]

The recent advent of large language models (LLMs) has introduced probabilistic priors that can be fused with neural evidence in BCI decoding pipelines.[4] By leveraging linguistic context learned from massive corpora, researchers have begun integrating LLMs into closed-loop BCI communication systems to reduce keystrokes and accelerate sentence composition. Early work in BCI spelling incorporated relatively simple n-gram language models as priors to improve character/word prediction and typing efficiency;[5] more recent systems extend this idea using LLMs for context-aware word suggestions and higher-level assistance.[6] These LLM-enabled interventions can plausibly span intent expansion, post-editing, and adaptive interface behaviors, with the broader assistive-communication literature already demonstrating LLM-driven UI strategies that

reduce user effort.[7]

Despite the rapid proliferation of proof-of-concept LLM architectures, there remains a critical need to systematically evaluate how these integrations alter the performance landscape of communication BCIs. The objective of this systematic review is to comprehensively map the taxonomy of LLM integration into BCIs, evaluate the methodological rigor of current literature, and quantify the resultant impacts on BCI efficacy. We explicitly evaluate the safety constraints, API dependencies, and real-time decoding challenges present in the current generation of LLM-assisted BCIs.

II. Methods

A. Protocol and Registration

This study used an established methodological framework, consistent with the extended Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist for systematic reviews.[8] The study protocol was pre-registered on the International Prospective Register Reviews (PROSPERO; CRD420261321575).

B. Eligibility Criteria

Studies were evaluated based on the PICOS framework (Population, Intervention, Comparator, Outcomes, and Study design). Studies were included if they met all of the following criteria: (1) involved a population of individuals utilizing non-invasive or invasive BCIs for communication or control, including both healthy volunteer cohorts and target patient populations; (2) applied an intervention consisting of the integration of a deep learning-based Large Language Model (e.g., GPT, LLaMA, BERT) directly within the BCI processing pipeline or user interface loop; (3) reported a comparator against classical BCI spelling interfaces (e.g., standard letter-by-letter spellers) lacking LLM enhancement, or baseline n-gram models where applicable; (4) reported objective metrics of BCI efficacy including spelling accuracy, typing speed, Information Transfer Rate (ITR), Keystroke Savings (KS), or qualitative usability measures; and (5) consisted of primary research articles evaluating offline simulations or online real-time implementations.

C. Data Sources and Searches

We conducted a comprehensive literature search across three academic databases: PubMed, IEEE Xplore, and arXiv. The search strategy combined free-text keywords across two conceptual domains: (1) brain-computer interfaces (”Brain-Computer Interface” OR ”BCI” OR ”Neural Interface”) and (2) large language models (”Large Language Model” OR ”LLM” OR ”ChatGPT” OR ”Generative AI”). The full search strategy for each database is provided in the Supplementary Methods. We additionally performed forward and backward citation searching of included studies to identify further relevant articles.

D. Data Extraction and Synthesis

We developed a standardized data extraction form to capture study-level characteristics and experiment-level technical specifications. Extracted variables included: clinical demographics, BCI modality, evaluation paradigm, LLM version, prompting strategy, API integration methods, and latency profiles. Two reviewers (A.G., M.O.) independently extracted data from each included study, with any discrepancies resolved through third reviewer (E.K.). Finally, risk of bias was evaluated utilizing an adapted Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) framework, which is appropriate for the single-arm and crossover experimental designs predominant in BCI literature. Key evaluated domains included baseline confounding, measurement of interventions, and selection of reported results. Certainty of evidence was appraised narratively following GRADE principles.

Given the substantial heterogeneity in BCI paradigms, target populations, and integration methodologies across the included studies, we did not conduct a quantitative meta-analysis. Instead, we performed a narrative synthesis supported by a tabular taxonomy, structured around our primary outcomes of effective communication rates and keystroke savings.

E. Risk of Bias and Certainty of Evidence

Risk of bias was evaluated utilizing an adapted Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) framework, appropriate for the single-arm and crossover experimental designs predominant in BCI literature. Key domains included baseline confounding, measurement of interventions, and selection of reported results. Certainty of evidence was appraised narratively following GRADE principles, specifically prioritizing the distinction between offline simulations and online real-time validations.

F. Standardized BCI Metric Definitions

To ensure consistency across disparate studies, the following standard BCI metrics and equations were strictly defined:

Accuracy (P) was defined as the ratio of correct selections to total selections:

$$Accuracy(P) = \frac{\text{Number of correct selections}}{\text{Total selections}} \tag{1}$$

Information Transfer Rate (ITR) was calculated utilizing the standard Wolpaw definition, where N represents the number of targets, P denotes selection accuracy, and T signifies the time required per selection (seconds):

$$ITR = \left[\log_2(N) + P\log_2(P) + (1-P)\log_2\frac{1-P}{N-1}\right]\cdot\frac{60}{T} \tag{2}$$

Keystroke Savings (KS) quantified the reduction in user effort necessitated by the LLM predictive capabilities:

$$KS = 1 - \frac{K_{\text{assisted}}}{K_{\text{baseline}}} \tag{3}$$

where K_{assisted} and K_{baseline} represent the number of keystrokes executed in the LLM-enhanced condition and the standard control condition, respectively.

III. Results

A. Study Characteristics and PRISMA Flow

Database searches and manual citation tracking initially identified 1049 unique records. Following abstract and title screening, 55 full-text articles were acquired for rigorous assessment. Ultimately, nine studies met the comprehensive inclusion criteria (**Figure 1.**).[9–17]

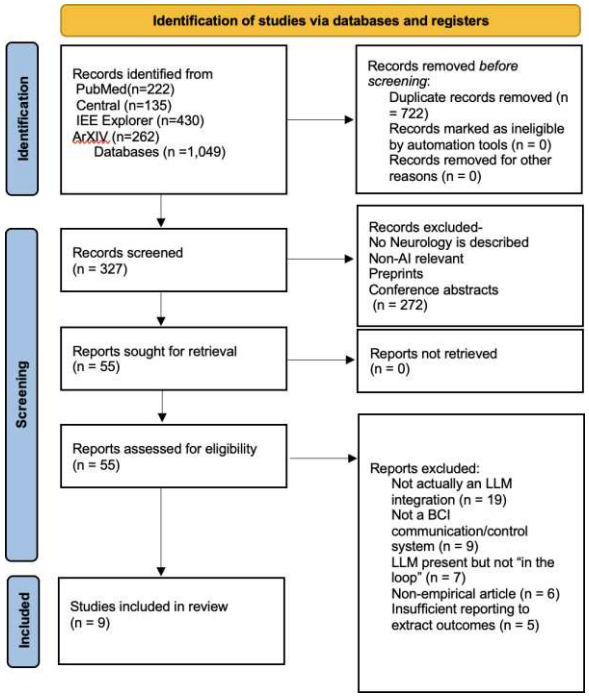


Fig. 1. **Figure 1. PRISMA Flow Diagram**

The included literature comprises evaluations of 46 healthy participants and a simulation cohort derived from previously collected datasets. The paradigms spanned visually evoked P300 spellers ($n = 3$), Code-Modulated Visual Evoked Potentials (cVEP; $n = 1$), Steady-State Visual Evoked Potentials (SSVEP; $n = 2$), and passive affective computing pipelines ($n = 1$), alongside auditory feasibility tests ($n = 1$). The majority of the studies utilized online real-time decoding, underscoring a maturity beyond purely offline simulations.

B. BCI and LLM Pipeline Integration

Traditional deterministic pipelines have been substantially augmented by large-parameter models.

The LLM pipelines relied heavily on remote API calls (primarily OpenAI GPT-3.5 and GPT-4o), operating predominantly in a zero-shot capacity with strict system prompts enforcing intent compression and structured output formatting. Safety constraints varied, with some studies utilizing fine-tuning to mitigate factual hallucinations in sentence expansion tasks.

C. Taxonomy of LLM Integration Patterns

We identify five dominant integration paradigms for BCI-LLMs:

1) **Autocomplete and Next-Word Prediction:** Utilizing contextual dialogue history to offer 3–10 dynamic word completions during spelling (e.g., Cantu¨rk & Volosyak, 2024; Hong et al., 2025).[10,13]
2) **Post-Edit Correction:** Feeding noisy, misclassified string sequences into an LLM to probabilistically reconstruct the intended spelling without requiring additional user trials (e.g., Lebedev et al., 2025).[14]
3) **Intent Expansion:** Mapping singular BCI-selected keywords into coherent, contextually appropriate sentences (e.g., Meunier et al., 2024).[9]
4) **Dynamic UI Generation:** Allowing the LLM to dynamically construct the BCI stimulus matrix (e.g., determining the frequencies and layouts for SSVEP) based on inferred user goals (e.g., Jin et al., 2025).[11]
5) **Affective Support:** Interpreting passive EEG emotional valence to adapt AI conversational agents (e.g., Sorino et al., 2024).[12]

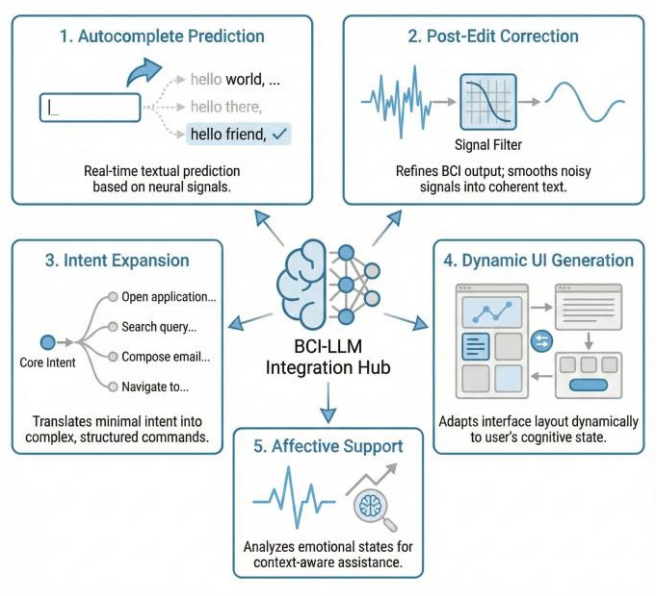


Fig. 2. **Figure 2. Neuro-computational Taxonomy of BCI-LLM Integration.** This schematic defines the five primary functional integration modalities where Large Language Models augment Brain-Computer Interfaces. The central hub illustrates the core neuro-symbolic bridge: translating biological waveforms into discrete transformer architectures. Symmetrically branching outwards are the five distinct algorithmic functions: (1) *Autocomplete Prediction*: generating high-dimensional predictive sequences from partial neural inputs; (2) *Post-Edit Correction*: mapping noisy, error-prone classification signals onto probabilistically aligned semantic manifolds; (3) *Intent Expansion*: employing deep knowledge graphs to algorithmically expand low-bandwidth seed commands into dense, complex syntactic structures; (4) *Dynamic UI Generation*: performing autonomous real-time state-space optimization to fluidly reconfigure BCI interface layouts based on user constraints; and (5) *Affective Support*: synthesizing multimodal physiological data (e.g., EEG valence arrays) to dynamically alter dialogue feedback loops and agentic triage.

TABLE I
TABLE 1. INCLUDED STUDIES AND DESIGN

Study (year), venue	Country	Design	N (participants) and population	BCI modality + paradigm	LLM model(s)	Primary integration pattern	Comparator(s)
Hong et al. (2025)_Scientific Reports	USA	Online real-time	N=7Healthy	EEGP300	GPT-3.5-turbo	Autocomplete (Word completion & prediction)	Letter-by-letter P300 speller
Wang et al. (2025)_arXiv_	China	Online real-time	N=10Healthy	EEGSSVEP	GPT-4o(GPT-4o-mini tested offline)	Autocomplete (Next word/sentence prediction)	Naive BCI (no LM); Directed Word Graph (DWG)
Meunier et al. (2024)_arXiv_	Austria	Online real-time	N=5Healthy	EEGc-VEP	GPT-3.5 (keywords)GPT-3 Davinci (sentences)	Intent expansion (Keyword to full sentence)	None (Proof of concept)
Parthasarathy et al. (2024) _arXiv_	USA	Offline simulation	N=78 (dataset)Healthy	EEGP300	GPT-2, BERT, BART, XLNet, RoBERTa, GPT-Neo	Autocomplete (Next word prediction)	Ran-dom/Deterministic flashboards; Dijkstra algorithm
Cantu¨rk & Volosyak (2024) _9th BCI Conf	Germany	Online real-time	N=13Healthy	EEGc-VEP	ChatGPT (gpt-3.5-turbo)	Autocomplete (Word prediction)	Historical n-gram model control
Lebedev et al. (2025)_Preprint _	Russia	Simulation (using real data)	N=10 (dataset)Healthy	EEGP300	ChatGPT, DeepSeek, Grok	Correction (Post-editing noisy character streams)	Varying repetition counts (2, 5, 10 flashes)
Jin et al. (2025) _arXiv _	China	System Proposal	Not reported (system architecture/proof-of-concept; no user study details given)	EEGSSVEP	Generic "Large Model"	Interface generation (Dynamic layout & command reasoning)	-
Sorino et al. (2024)_UMAP Adjunct_	Italy	Framework Proposal	N=30 (training)Healthy	EEGPassive (Emotion)	LLaMA 2 (chat)	Affect support (Emotional support conversation)	-
Žak & Grosse-Wentrup (2024)_Preprint_	Austria	Feasibility Study	N=10Healthy	AuditoryReactive (AID)	None (Concept extends EEGChat)	Intent expansion (Conceptual; study focused on decoding)	Chance level

D. Quantitative Outcomes

Performance gains were broadly consistent across the interventions. Where baseline classical spellers were used as comparators, LLM assistance yielded striking enhancements in time-to-completion, driving Information Transfer Rates significantly higher by increasing the practical bits-per-selection metric over raw channel capacity. Keystroke Savings (KS) consistently exceeded 50%, peaking above 80% in highly constrained scenario-based tasks.

E. Risk of Bias

According to the adapted ROBINS-I criteria, the methodological quality was generally acceptable for early-stage prototype evaluation. However, domain-level bias emerged concerning the selection of participants (overwhelmingly reliant on convenience cohorts of healthy, neurotypical university students) and missing data (limited reporting of end-to-end latency and API failure rates).

IV. DISCUSSION

In our review, we found that LLM-assisted BCI studies have started to accumulate rapidly in the recent literature. The synthesized results showcase that LLM integration consistently enhances effective Communication Rates and significantly reduces user fatigue via Keystroke Savings across multiple BCI paradigms (e.g., P300, SSVEP, and cVEP). Across these implementations, we noticed certain structural patterns which we identify in a novel 5-part integration taxonomy: Autocomplete, Post-Edit Correction, Intent Expansion, Dynamic UI Generation, and Affective Support. However, a synthesis of this current evidence base mandates a continuing separation between rigorously proven capabilities and theoretical speculation.

A. Proven vs. Speculative Capabilities

What is currently proven across multiple cohorts (both online and simulated) is that LLM assistance fundamentally reduces the sheer quantity of electrophysiological selections required to convey an intent. This is most vividly demonstrated by Keystroke Savings (often exceeding 50%) and concurrent increases in the effective ITR of classical P300 and cVEP paradigms. What remains highly speculative is the clinical safety and long-term viability of dynamic, unconstrained intent-expansion systems (e.g., generating entire paragraphs from a single keyword) in a population lacking communication, given the propensity of LLMs to hallucinate extraneous or subtly incorrect context. Recent studies emphasize that the

TABLE II
TABLE 2. BCI PIPELINE DETAILS

Study	Acquisition	Paradigm parameters	Preprocessing / features	Decoder / classifier	Calibration & adaptation	Interface / UI characteristics
Hong et al. (2025) _ChatBCI _	Device: Brain Products actiCHamp PlusCh: 16 (of 32)Rate: 250 Hz	Type: P300 (Row-Column)Stim: 40 ms flash, 100 ms ISIReps: 8 repetitions	Filter: 0.5-30 HzFeatures: Moving average (decimation factor 12), window 0-700 ms	SWLDA (Stepwise Linear Discriminant Analysis)	Calibration: 20 specific keys, 5 repetitions each; personalized model trained	Matrix: 5x8 (40 keys)Layout: 26 letters + function keys + 10 dynamic LLM suggestion slots
Wang et al. (2025) _MindChat _	Device: g.USBampCh: 9 (Pari-etal/Occipital)Rate: 1200 Hz	Type: SSVEPStim: 1.5s duration; Freq 8-16 Hz (0.2 Hz step), Phase 0-1.5π	Filter: 7-70 HzResample: 240 Hz	FBECCA (Filter-bank extended canonical correlation analysis)	Template: Reference signals derived from sine/cosine and multiple SSVEP trials	Matrix: 40 keysLayout: Virtual keyboard with 7 candidate index keys for LLM predictions
Meunier et al. (2024) _EEGChat _	Device: Bittium NeurOne TeslaCh: 6 (Occipi-tal/Parietal)Rate: 1000 Hz	Type: c-VEP (m-sequence)Stim: 20 Hz flash rate, length 31Reps: Max 7 cycles (10.85 s)	Window: 250 msFeatures: Unfiltered EEG fed to CNN	EEG2Code (CNN) + Correlation (Sequence matching)	Training: Minimum 7 interactions (1519 frames) to train decoder	Layout: 6 keyword options + 4 special options (Correction, More, None, Finished)
Parthasarathy et al. (2024)	Device: GTEC amplifiersCh: 32Rate: 256 Hz	Type: P300 (Row-Column)Stim: 100 ms flash, 25 ms ISIMatrix: 6x6	Filter: 0.1-60 HzFeatures: Downsampled by factor 12	SWLDA (Stepwise Linear Discriminant Analysis)	Training: Tested both Across-Subject (ASCV) and Within-Subject (WSCV) cross-validation	Layout: Virtual flashboards (Diagonal, Frequency sorted) overlayed on standard 6x6 grid
Cantu¨rk & Volosyak (2024)	Device: g.USBampCh: 16Monitor: 240 Hz	Type: c-VEP (m-sequence)Stim: 2.1s trial (2 cycles, 63-bit code)Targets: 8	Method: CCA spatial filters developed during recording	CCA (Canonical Correlation Analysis)	Calibration: 6 blocks of 8 stimuli (48 trials total)	Layout: "Two-step speller"; 8 boxes (2x4); Top row letter groups, Bottom row ChatGPT suggestions
Lebedev et al. (2025)	Device: NVX-36 amplifierCh: 3 (P3, Pz, P4)Rate: 250 Hz	Type: P300Stim: 4x4 grid, 10 flashes/charSim: Tested 2, 5, 10 reps	Preprocessing: Not reported (in this paper). Simulation pools on-target vs. non-target trials; correlation coefficient r computed between average EEG trace and target template (template from average of 100 target responses).	Correlation coefficient (r) matching against target template	Template: Pooled epochs from on-target vs. non-target trials	Matrix: 4x4 grid (16 chars)Visual: Color-coded r-value matrix feedback
Jin et al. (2025)	Device: Wireless EEGCh: 64Rate: 1000 Hz	Type: SSVEPStim: Dynamic blinking blocks	Filter: 50Hz notch, artifact removalMethod: TDCA enhancement	TDCA (Task-Discriminant Component Analysis)	Calibration: Individual calibration using time-delayed data augmentation	Dynamic UI: LLM generates layout, frequency, and text based on semantic context (e.g., home control)
Sorino et al. (2024) _ARIEL _	Device: Muse 2 (EEG)Ch: Not reportedRate: Not reported	Type: Passive BCI (Emotion)Stim: Music videos (for training)	Filter: 1-40 Hz bandpassFeatures: Artifact removal (meegkit)	SVM (Support Vector Machine) outperformed RF, XGB, LR	Training: Leave-One-Subject-Out (LOSO) on 30 subjects	UI: Chatbot interface; BCI operates passively to modulate LLM persona
Ź ak & Grosse-Wentrup (2024)	Device: Not specifiedCh: Not specified	Type: Auditory Intention Decoding (AID) / N400 ERPStim: 3-digit numbers	Window: -0.2 to 1.85s relative to stimulus	EEGNet (CNN)	Training: 10-fold cross-validation	Auditory UI: System reads options; user attends to target option

transition from assistive LLMs to autonomous models such as BCI-LLM exacerbates these vulnerabilities.[18–20]

B. Mechanisms of LLM Enhancement

LLM assistance transcends the limitations of classic *n*-gram models primarily through robust language priors, which enable deeper intent compression.[21] While traditional spellers combat BCI decoding noise via repetitive trial averaging (increasing dwell time)[22], LLMs approach the noisy channel problem via contextual error correction. As shown by Lebedev et al., highly parameterized models can successfully recover the intended spelling even when the raw BCI character decoding accuracy is structurally impaired by reduced trial repetitions.[14] While classical n-gram language models operate nearly instantaneously as local probability look-up tables, the forward passes of highly parameterized LLMs (often comprising billions of parameters) mandate massive matrix multiplications for every generated token. This architectural necessity severely impacts both Time To First Token (TTFT) and Inter-Token Latency. Consequently, reliance on edge-to-cloud telemetry to offload these computations introduces stochastic network jitter into synchronous BCI polling loops.[23] For online control, delays exceeding a few hundred milliseconds risk stalling cognitive

TABLE III
TABLE 3. LLM PIPELINE AND HOW IT IS USED

Study	LLM (name, version)	Local vs API	Prompting	LLM inputs	LLM outputs	Safety constraints	Latency + failure handling
Hong et al. (2025)_ChatBCI _	GPT-3.5-turbo	API (OpenAI)	Zero-shotRole-taking (System/User); Temp=0 for deterministic behavior.	Partial text composed by user.	Suggestions list: 10 candidate words (completion or prediction).	Refusal rulesPrompt fine-tuned to ensure valid suggestion format; strictly formatted response requested.	Latency: Implicit in task time; API cost analyzed ($0.05/100 words).
Wang et al. (2025)_MindChat _	GPT-4o(gpt-4o-2024-08-06)_Tested: GPT-4o-mini _	API (OpenAI)	Few-shotSystem prompt + 3 context examples (e.g., prefix completion).	Current text input + Dialogue context (history).	Suggestions list: JSON format: Prefix + 5 words + 2 sentences.	Restricted outputStrict JSON format required; no extra description allowed.	Latency: 1s-2s average response time.Failure: Users can fallback to letter-by-letter typing.
Meunier et al. (2024)_EEGChat _	GPT-3.5 (Keywords)GPT-3 Davinci (Sentences)	API	MixedKeywords: Prompt-engineered (Temp 0.5).Sentences: Fine-tuned.	Transcribed partner question + Conversation history.	Intent expansion: 6 Keywords -¿ Full spoken sentence.	Restricted outputFine-tuned (FT-HQ) to reduce hallucinations (”Added details”) and ”Wrong info”. UI includes ”Correction” and ”None”.	Latency: Not reported.Failure: Audio transcription retries if failed.
Parthasarathy et al. (2024)	GPT-2, BERT, BART, XLNet, RoBERTa, GPT-Neo	Local (open-source LMs; offline simulation)	N/AStandard probability estimation / Next-token prediction.	Context of preceding words/characters.	Auto-complete: N-best list; Layered with Dijkstra for OOV fallback.	N/A(Offline simulation).	-
Cantu¨rk & Volosyak (2024)	ChatGPT(gpt-3.5-turbo-0125)	API	System prompt”Always return suggestions for just the most likely last word...”	Typed sentence/query.	Suggestions list: 3 comma-separated words.	Restricted outputPenalties (Freq/Pres = -0.1) to reduce randomness; Top_p = 0.1 for fo-cus.max_tokens=100	Latency: Occasional unresponsive-ness/delays due to network traffic.Failure: Sometimes output missing parts of words despite instructions.
Lebedev et al. (2025)	ChatGPT, DeepSeek, Grok	Not specified (likely Web/API)	Zero-shot”Correct the spelling” instruction.	Noisy character sequence (e.g., ”HT WASHED HIS HAN S...”).	Correction: Reconstructed sentence.	N/A(Simulation).	-
Jin et al. (2025)	LLM not specified (invokes an LLM API)	API (HTTP request)	Not specified (LLM parses intent and returns interface/control specification)	Natural language command (e.g., ”fly the drone...”).	Commands & UIJSON with device status/function + Dynamic UI parameters (layout, frequency).	Restricted outputConversion to specific protocols (e.g., $Robot arm(10, 20, 30)).	Latency: Not reported.Failure: Not reported.
Sorino et al. (2024)_ARIEL _	LLaMA 2 (chat)	Local / Hosted	Dynamic Hard-PromptingRole-play prompts updated based on conversation state.	User message + BCI Emotion Label.	Affect support-Empathetic natural language response.	ConstraintPrompt evolution ensures relevance to emotional state.	-
Žak & Grosse-Wentrup (2024)_Preprint_	Not applicable (LLM integration proposed as extension to EEGChat).	Not applicable.	Not applicable.	Not applicable.	Not applicable.	Not applicable.	Not reported.

TABLE IV
TABLE 4. OUTCOMES AND EVALUATION

Study	Task type	Accuracy	Speed	ITR (bits/min)	Utility	User-reported outcomes	Adverse events / Fatigue
Hong et al. (2025)_ChatBCI _	1. Copy spelling: Self-composed sentence based on image2. Improvised: Free spelling	Selection Acc: 88.77% (ChatBCI) vs 77.89% (Control)Char Acc: 100% (ChatBCI) vs 94.64% (Control),	Time: 10.43 min (ChatBCI) vs 27.55 min (Control)Rate: 2.49 char/min (ChatBCI) vs 0.94 (Control)	Copy: 26.12 (Realistic estimate)Improv: 73.95 (Realistic estimate)Formula: $N=30+M$ (avg chars in suggestions),	Keystroke Savings (KS):53.22% (Copy)80.68% (Improvised)KS-Deficit Ratio: 30% (Copy), 2.28% (Improv),	Inferred Usability: 2 subjects failed to complete the task in Control (letter-by-letter) mode but succeeded with ChatBCI	None explicitly reported, though reduced time implies reduced fatigue
Wang et al. (2025)_MindChat _	**Copy spelling:**Single-turn & Multi-turn dialogues (Daily & Healthcare)	Decoding Acc: ˜89% (FBECCA)Focus was on system efficiency, not accuracy per se	Time: Reduced ¿30% compared to naive spellerSpeedup: Multi-turn dialogues showed higher speedup rates	Not reported	Keystroke Savings: ¿62% overallMulti-turn saved more keystrokes (64%) than single-turn (55%),	Strategy: Users adopted ”shortcut strategies” (selecting similar sentence + delete word) to speed up typing	None reported
Meunier et al. (2024)_EEGChat _	Dialogue: Simulated goal-oriented phone calls (e.g., booking appointment)	Task Success: 4/5 subjects accomplished all goals; S5 failed 2/6 tasks	Selection Time: 2.23s to 9.69s (variable by subject)Completions: Avg 8.68 utterances per task	Not reported	Preference: First two keywords selected 48% of the time (LLM prioritized likely tokens)	**Survey:**Understanding goals: 5/5Keywords relevant: 5/5Answers matched intent: 4/5 (S2 disagreed)	Fatigue/Frustration: S5 failed tasks due to misclassifications; often reached max stimulation time (10.85s),
Parthasarathy et al. (2024)	Offline Simulation: Copy spelling ”Declaration of Independence”	Retry Rate: Near 0% for GPT-2/Diagonal models (vs significant backspacing in Random)	Speedup: 10% char-level40% multi-word prediction	**Within-Subject: 75.31 (Diagonal+GPT2) Across-Subject:**59.75 (Diagonal+GPT2)Formula: Standard log formula	Utility: OOV (Out of Vocabulary) handling improved via smoothing	N/A (Simulation)	-
Cantu˘rk & Volosyak (2024)	**Copy spelling:**1. Words (BCI, KLEVE)2. Pangram3. Unique sentence	Accuracy: 87.49% (Mean)	**OCM (Output Chars/Min):**8.74 (Unique sentence)17.72 (Pangram)Slower than n-gram baseline (18.9),	Mean: 33.16 bits/min	Acceptance: 12/13 found dictionary hints helpful	**Survey:**13/13 preferred ChatGPT aid12/13 would repeat experiment	Fatigue: 8/13 felt ”more tired” after exp.Visual Strain: 4/13 found flickering disturbing
Lebedev et al. (2025)	Simulation: Copy spelling words (”FISH”) and sentences	Decoding Acc:40% (2 reps) $\to$ 100% (10 reps)Sentence Recovery: 100% via LLM	Not reported (Concept: ”Speedy but imprecise”)	Not reported	Correction: LLM successfully reconstructed intended sentences from noisy input strings	N/A (Simulation)	-
Jin et al. (2025)	Command/Control: Robot arm, Drone, Home (System Proposal)	Not reported	Not reported	Not reported	Translation: Natural language commands converted to protocol codes (e.g., $Robot arm(10, 20, 30))	Not reported	-
Sorino et al. (2024)_ARIEL _	Passive BCI: Emotion Recognition during chat	**Emotion Classification:**˜0.77-0.80 (Happy, Angry, Sad, Relaxed)	-	-	-	Running Example: Demonstrated successful prompt adaptation based on detected emotion	-
Z̆ak & Grosse-Wentrup (2024)	Feasibility: Auditory selection of numbers	Decoding Acc: 0.6 - 0.7 (Binary, above chance)	Not reported	Not reported	-	-	-

TABLE V
TABLE 5. RISK OF BIAS JUDGMENTS USING ADAPTED ROBINS-I

Study	Baseline Confounding	Selection of Participants	Classification of Interventions	Deviations from Interventions	Missing Data	Measurement of Outcomes	Selection of Reported Result	Overall Risk of Bias
Hong et al. (2025)	Low	Moderate	Low	Low	Low	Low	Low	Low
Wang et al. (2025)	Low	Moderate	Low	Low	Low	Low	Low	Low
Meunier et al. (2024)	Low	Moderate	Low	Low	Moderate	Low	Low	Moderate
Parthasarathy et al. (2024)	Low	Low	Low	Low	Low	Low	Low	Low
Cantu¨rk & Volosyak (2024)	Low	Moderate	Low	Low	Low	Low	Low	Low
Lebedev et al. (2025)	Low	Moderate	Low	Low	Low	Low	Low	Low
Jin et al. (2025)	Moderate	High	Moderate	Low	High	High	High	High
Sorino et al. (2024)	Moderate	High	Low	Low	High	Low	High	Moderate
Z˘ a´k & Grosse-Wentrup (2024)	Low	High	Low	Low	High	Low	Low	Moderate

state queues, actively desynchronizing the patient from the interface, and disrupting the precise rhythmic attention required by traditional paradigms.[24]

C. The Offline vs. Online Gap and Clinical Validation

A substantial methodological gap separates offline simulation from real-time implementation. Offline simulations inherently ignore the interactive cognitive dynamics of the BCI user adjusting to the pacing of the LLM outputs.[25–27] True clinical validation demands online deployment in motor-impaired patient cohorts a glaring deficiency in the current literature, which relies almost entirely on neurotypical university students. Validating BCI performance on healthy cohorts may misrepresent target clinical deployments. Patients with advanced neuromuscular diseases, such as amyotrophic lateral sclerosis (ALS) or locked-in syndrome (LIS), exhibit significantly altered cortical dynamics, reduced signal-to-noise ratios, and varying cognitive states compared to intact neurotypical models.[28] It remains highly unproven whether LLMs can accurately infer complex, multi-sentence contexts from the faint, non-stationary single-thought triggers characteristic of completely locked-in states.[29]

D. Recommendations for Standardized Reporting

To mature the field, we vehemently urge the adoption of standard reporting frameworks. Studies must provide minimum dataset descriptions detailing clinical demographics, exact BCI impairment scales (for patient cohorts), and full hardware montages. We enforce strict adherence to standardized ITR calculation alongside unmodified raw selection accuracies. Furthermore, researchers must separately report biological dwell-time, decoding network execution time, and LLM query-to-response latency. Finally, mandatory ablation baselines including both a no-LM baseline and a classic local n-gram baseline are required to rigorously justify the computational overhead of the LLM.

E. Research Agenda

Next steps should pivot rapidly toward *local*, constrained, and on-device decoding using smaller models customized via Low-Rank Adaptation (LoRA) to preserve privacy and minimize latency. Furthermore, longitudinal home-deployment studies involving active clinical patient cohorts, particularly those studying multilingual support and pervasive personalization, are the required benchmark for the next evolutionary stage of this technology.

V. LIMITATIONS

This systematic review is constrained by the recency state of the literature. The small sample sizes, heterogeneity in evaluation paradigms, and near absence of disabled target populations restrict the generalizability of the findings and precluded quantitative meta-analysis.

VI. CONCLUSION

The integration of LLMs into BCI communication frameworks delivers empirically substantive improvements to transfer rates and limits user fatigue. As the BCI architecture shifts from a deterministic signal processing pipe to an interactive human-AI collaborative dialogue, overcoming the persistent hurdles of latency, neuro-privacy, and AI reliability will distinguish theoretical novelties from life-altering clinical prosthetics.

ACKNOWLEDGMENTS

The authors acknowledge the anonymous reviewers and the open science community for establishing rigorous evaluation paradigms in neurotechnology.

REFERENCES

[1] Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol*. 2002;113(6):767-791. doi:10.1016/s1388-2457(02)00057-3

[2] Chen X, Wang Y, Nakanishi M, Gao X, Jung TP, Gao S. High-speed spelling with a noninvasive brain–computer interface. *Proc Natl Acad Sci U S A*. 2015;112(44):E6058-E6067. doi:10.1073/pnas.1508080112

[3] Felton EA, Williams JC, Vanderheiden GC, Radwin RG. Mental Workload during Brain-Computer Interface Training. *Ergonomics*. 2012;55(5):526-537. doi:10.1080/00140139.2012.662526

[4] Speier W, Arnold C, Pouratian N. Integrating Language Models into Classifiers for BCI Communication: A Review. *J Neural Eng*. 2016;13(3):031002. doi:10.1088/1741-2560/13/3/031002

[5] Kindermans P jan, Verschore H, Verstraeten D, Schrauwen B. A P300 BCI for the Masses: Prior Information Enables Instant Unsupervised Spelling. In: *Advances in Neural Information Processing Systems*. Vol 25. Curran Associates, Inc.; 2012.

[6] Brain-to-Text Decoding with Context-Aware Neural Representations and Large Language Models. Accessed February 22, 2026. https://arxiv.org/html/2411.10657v1

[7] Zhao Y, Chen Y, Cheng K, Huang W. Artificial intelligence based multimodal language decoding from brain activity: A review. *Brain Res Bull*. 2023;201:110713. doi:10.1016/j.brainresbull.2023.110713

[8] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Published online March 29, 2021. doi:10.1136/bmj.n71

[9] Meunier A, Z˘a´k MR, Munz L, et al. A Conversational Brain-Artificial Intelligence Interface. *arXiv*. Preprint posted online March 14, 2024:arXiv:2402.15011. doi:10.48550/arXiv.2402.15011

[10] Cantu¨rk, Atilla, Volosyak, Ivan. A NOVEL CHATGPT-DRIVEN COMMUNICATION AID BASED ON CODE-MODULATED VISUAL EVOKED POTENTIALS (CVEP). doi:10.3217/978-3-99161-014-4-008

[11] Jin J, Zhang Y, Xu R, Chen Y. An Innovative Brain-Computer Interface Interaction System Based on the Large Language Model.

[12] Sorino P, Biancofiore GM, Lofu` D, et al. ARIEL: Brain-Computer Interfaces meet Large Language Models for Emotional Support Conversation. In: *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. ACM; 2024:601-609. doi:10.1145/3631700.3665193

[13] Hong J, Wang W, Najafizadeh L. ChatBCI, a P300 speller BCI with context-driven word prediction leveraging large language models, from concept to evaluation. *Sci Rep*. 2026;16(1):6379. doi:10.1038/s41598-025-25660-7

[14] Lebedev MA, Makarova AV, Kleeva DF, Maysuradze AI. Connecting a P300 speller to a large language model. *bioRxiv*. Preprint posted online November 7, 2025:2025.11.06.686984. doi:10.1101/2025.11.06.686984

[15] Parthasarathy N, Soetedjo J, Panchavati S, et al. Evaluation Of P300 Speller Performance Using Large Language Models Along With Cross-Subject Training. *arXiv*. Preprint posted online October 19, 2024:arXiv:2410.15161. doi:10.48550/arXiv.2410.15161

[16] Wang Ji, Zhong Y, Huang C, Yao L. MindChat: Enhancing BCI Spelling with Large Language Models in Realistic Scenarios. *arXiv*. Preprint posted online July 29, 2025:arXiv:2507.21435. doi:10.48550/arXiv.2507.21435

[17] Z˘ a´k MR, Grosse-Wentrup M. TITLE: AUDITORY CONVERSATIONAL BAI: A FEASIBILITY STUDY.

[18] Gorenshtein A, Omar M, Glicksberg BS, Nadkarni GN, Klang E. AI Agents in Clinical Medicine: A Systematic Review. *medRxiv*. Published online August 26, 2025:2025.08.22.25334232. doi:10.1101/2025.08.22.25334232

[19] Klang E, Glicksberg BS, Gorenshtein A, et al. Clinical Agents Don't Care. *medRxiv*. Preprint posted online October 19, 2025:2025.10.17.25338226. doi:10.1101/2025.10.17.25338226

[20] Omar M, Gorenshtein A, Agbareia R, et al. Impact of patient communication style on agentic AI-generated clinical advice in E-medicine. *Am J Med*. Published online January 20, 2026:S0002-9343(26)00010-0. doi:10.1016/j.amjmed.2025.12.027

[21] Ghaseminejad Raeini M. The evolution of language models: From N-Grams to LLMs, and beyond. *Nat Lang Process J*. 2025;12:100168. doi:10.1016/j.nlp.2025.100168

[22] Lu J, Speier W, Hu X, Pouratian N. The Effects of Stimulus Timing Features on P300 Speller Performance. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol*. 2013;124(2):306-314. doi:10.1016/j.clinph.2012.08.002

[23] Willett FR, Kunz EM, Fan C, et al. A high-performance speech neuroprosthesis. *Nature*. 2023;620(7976):1031-1036. doi:10.1038/s41586-023-06377-x

[24] Metzger SL, Littlejohn KT, Silva AB, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*. 2023;620(7976):1037-1046. doi:10.1038/s41586-023-06443-4

[25] Mattout J, Perrin M, Bertrand O, Maby E. Improving BCI performance through co-adaptation: Applications to the P300-speller. *Ann Phys Rehabil Med*. 2015;58(1):23-28. doi:10.1016/j.rehab.2014.10.006

[26] Frontiers — Closed-loop motor imagery EEG simulation for brain-computer interfaces. Accessed February 22, 2026.

[27] Pan H, Ding P, Wang F, et al. Comprehensive evaluation methods for translating BCI into practical applications: usability, user satisfaction and usage of online BCI systems. *Front Hum Neurosci*. 2024;18:1429130. doi:10.3389/fnhum.2024.1429130

[28] Nijboer F, Sellers EW, Mellinger J, et al. A P300-based brain-computer interface for people with amyotrophic lateral sclerosis. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol*. 2008;119(8):1909-1916. doi:10.1016/j.clinph.2008.03.034

[29] Chaudhary U, Vlachos I, Zimmermann JB, et al. Spelling interface using intracortical signals in a completely locked-in patient enabled by auditory feedback. *Nat Commun*. 2022;13(1):1236. doi:10.1038/s41467-022-28859-8