WAVEANCHOR-AGENT: FROM WAVEFORM DETECTION TO INTERACTIVE EEG ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained electroencephalography (EEG) interpretation requires localizing waveform events and explaining their spatial and temporal relationships. We introduce WaveAnchor-Agent (WA-Agent), an evidence-grounded workflow for fine-grained, generative EEG interpretation, together with a rubric-based benchmark for evaluating generated descriptions and dialogue. We build a physician-annotated dataset covering nine waveform classes and train a multi-scale temporal detector to localize events by channel and time. A planner–responder architecture connects waveform detection, deterministic measurement, and visualization tools to natural-language responses, supporting interactive analysis grounded in explicit signal-level evidence. Our benchmark covers waveform detection, selection, description, multi-turn conversation, and visualization. Its description and conversation tasks use 1,550 manually reviewed atomic rubric criteria to assess factual content beyond lexical similarity. On our benchmark, the detector achieves 45.22% test mAP at temporal IoU 0.5. WA-Agent achieves 93.71% selection accuracy and LLM-judged partial-credit rubric scores of 68.73% and 55.52% for description and conversation, respectively. Together, the workflow and benchmark provide a foundation for generating and evaluating fine-grained EEG explanations grounded in localized waveform evidence.

1 INTRODUCTION

Clinical electroencephalography (EEG) interpretation requires identifying waveform types and understanding their distribution across electrodes and time (Kural et al., 2020). Epileptiform discharges, other abnormal activity, physiologic sleep waveforms, and artifacts can appear together within a short segment. Readers must identify the patterns present, their locations, and how their measurements relate to background activity. Public datasets such as TUAB and TUEV support recording- or segment-level abnormality and event classification (Lopez et al., 2015; Obeid & Picone, 2016). Pretrained EEG models such as LaBraM and CBraMod advance these tasks through transferable signal representations (Jiang et al., 2024; Wang et al., 2025). Such classification alone, however, provides limited support for locating and explaining waveform events across channels and time.

We approach fine-grained EEG interpretation as an interactive task: given an EEG segment and a natural-language request, the system answers questions about waveform types, locations, and measurements, and provides visualizations when requested. To support this task, we introduce WaveAnchor-Agent (WA-Agent), a tool-using framework that couples the pretrained WaveAnchor-Detector (WA-Detector) with a planner–responder architecture. WA-Detector localizes nine waveform classes by channel and time, using multi-scale temporal features and electrode interactions. Based on the request, the planner selects analysis and visualization tools, and the responder answers from their structured outputs. The agent can thus answer increasingly detailed questions about the same EEG segment.

To support training and evaluation, we build a nine-class waveform dataset from clinical EEG recordings and public corpora. Physician annotations specify the waveform type, electrode, onset, and offset of each event. WaveAnchor-Bench (WA-Bench), an AI-assisted, manually reviewed benchmark, presents matched EEG cases as selection, description, and conversation tasks, with a subset for visualization and tool sequencing. Using these data, we assess signal localization, factual answer quality, and tool use separately.

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

In these evaluations, WA-Detector achieves 45.22% test mAP at temporal IoU 0.5, and WA-Agent reaches 93.71% selection accuracy against silver references. Rubric scores and tool-use tests also support WA-Agent's ability to describe EEG segments, answer follow-up questions, and deliver requested plots. Model comparisons suggest that planning can use a smaller model while response generation benefits from a stronger backend.

Our contributions are threefold:

- We introduce WA-Bench, to our knowledge the first clinical EEG benchmark combining channel-level temporal annotations with case-specific atomic rubrics for waveform description and multi-turn dialogue. It pairs physician annotations of nine waveform classes with 1,550 manually reviewed criteria. The dataset and benchmark will be publicly released upon acceptance.
- We develop WA-Detector, a multi-scale temporal detector that localizes nine waveform classes by channel and time, using a backbone pretrained through self-supervised learning on approximately 1,170 hours of clinical EEG.
- We introduce WA-Agent, a planner–responder framework that grounds fine-grained EEG descriptions and multi-turn answers in channel- and time-localized waveform detections, quantitative measurements, and visualizations.

2 Related Work

Clinical EEG modeling. Public EEG corpora support clinical tasks ranging from seizure detection (Guttag, 2010; Shah et al., 2018) to abnormality and slowing assessment (Lopez et al., 2015; von Weltin et al., 2017). EEG modeling has expanded from compact convolutional classifiers (Lawhern et al., 2018) to self-supervised methods that learn transferable signal representations (Kostas et al., 2021; Jiang et al., 2024; Wang et al., 2025). Their downstream evaluations commonly focus on recording- or segment-level predictions. Dedicated interictal epileptiform discharge (IED) detectors move toward waveform-level analysis by classifying short epochs or annotated transients (Tjepkema-Cloostermans et al., 2018; Jing et al., 2020), but retain a binary IED/non-IED objective.

Multi-scale temporal detection. Channel- and time-specific waveform labels turn EEG analysis into a detection problem: predicting what occurs and where. This parallels image object detection, where YOLO jointly predicts classes and locations (Redmon et al., 2016), feature pyramids handle different scales (Lin et al., 2017a), and FCOS regresses boundaries without predefined anchors (Tian et al., 2019). ActionFormer applies multi-scale, anchor-free detection to video action intervals (Zhang et al., 2022), while InceptionTime captures patterns at different temporal scales (Ismail Fawaz et al., 2020). Applying temporal detection to EEG requires preserving waveform morphology and accounting for spatial patterns across electrodes.

EEG with large language models and agents. Recent work applies large language models to EEG analysis and interaction. NeuroLM connects EEG to language through text-aligned signal tokens (Jiang et al., 2025), while CerebraGloss trains a vision-language model on waveform images and detector-derived instruction data (Gu et al., 2026). Tool use (Yao et al., 2022) enables EEG workflow coordination and task delegation (Zhao et al., 2026; Zhou et al., 2026), pipeline generation (Wang et al., 2026), and deterministic validation of proposed analyses (Hou et al., 2026). These studies expand the role of language models from interpreting EEG inputs to directing and checking analyses.

3 Clinical EEG Data and WaveAnchor-Bench

3.1 Clinical EEG Data

We construct the primary dataset from clinical EEG recordings collected at a hospital. The supervised cohort contains 34 annotated training recordings and 26 test recordings, with no patient overlap between splits. From the test recordings, we extract 175 waveform-enriched 10-s segments. A separate collection of 770 long-term recordings (approximately 1,170 hours), with subjects disjoint from the supervised cohort, is used for self-supervised pretraining.

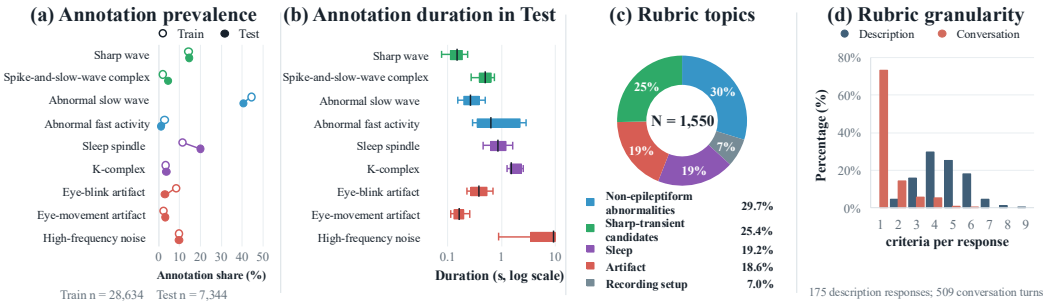


Figure 1: Clinical EEG dataset and WA-Bench statistics. (a) Within-split proportions of the nine channel-level annotation classes; open and filled markers denote training and test data, respectively. (b) Test-set annotation durations. Boxes show the interquartile range, black lines show the median, and whiskers show the 10th–90th percentiles on a logarithmic time axis. (c) Distribution of 1,550 atomic criteria across five rubric topics. (d) Distribution of the number of atomic criteria assigned to each description response or individual conversation turn.

All recordings are standardized to 19 channels from the international 10–20 system. We apply a 1.6–70 Hz band-pass filter and a 50 Hz notch filter, resample to 200 Hz, and use an average reference; amplitudes are represented in microvolts. Annotations were produced by an EEG specialist and subsequently reviewed by a medical trainee. Each annotation specifies waveform type, channel, onset, and offset, while sleep stage is retained from the source annotation. We use nine target categories: sharp wave, spike-and-slow-wave complex, abnormal slow wave, abnormal fast activity, sleep spindle, K-complex, eye-blink artifact, lateral eye-movement artifact, and high-frequency noise. The clinical subset contains 28,634 training events and 7,344 test events. The class distribution is long-tailed but broadly consistent across the training and test splits (Figure 1(a)). Annotation durations vary more than 100-fold (Figure 1(b)): sharp waves usually last under 1 s, whereas K-complexes and high-frequency noise typically last longer.

To increase waveform diversity, we include a small set of training examples from TUAB (Lopez et al., 2015), TUEV (Obeid & Picone, 2016), TUSL (von Weltin et al., 2017), and TUEP (Veloso et al., 2017). We annotate these examples under the same nine-class taxonomy; they are used only for training, while all evaluation data come from the held-out clinical cohort.

3.2 WaveAnchor-Bench Tasks

To evaluate WA-Agent, we build four tasks based on the 175 test EEG segments. **Selection** requires choosing the single correct answer from four options. **Description** requires a concise free-form account of waveform types, spatial distribution, timing, and artifacts. **Conversation** asks progressively focused questions about the same segment and evaluates response accuracy at each turn. **Waveform visualization** uses a 30-case subset with 60 interaction turns to test whether the agent correctly invokes and sequences waveform-rendering and analysis tools.

For description and conversation, reference answers are expressed as independently judgeable rubric criteria, such as "The description identifies sharp waves distributed bilaterally over frontal regions." A response receives credit for each criterion it satisfies. These rubric criteria cover five major semantic topics (Figure 1(c)). Description responses generally contain several independently scored criteria, whereas individual conversation turns usually target one criterion (Figure 1(d)). The visualization task is scored according to tool-routing accuracy. Detailed statistics and scoring protocols are provided in Appendices A and F, with worked rubric examples in Appendix C.4.

4 WaveAnchor-Detector

4.1 Problem Formulation

Let $\boldsymbol{X} \in \mathbb{R}^{C \times T}$ denote a 10-s EEG window with $C = 19$ channels and $T = 2000$ samples. Each target event is a tuple (c, s, e, y) containing an electrode index, normalized start and end times, and

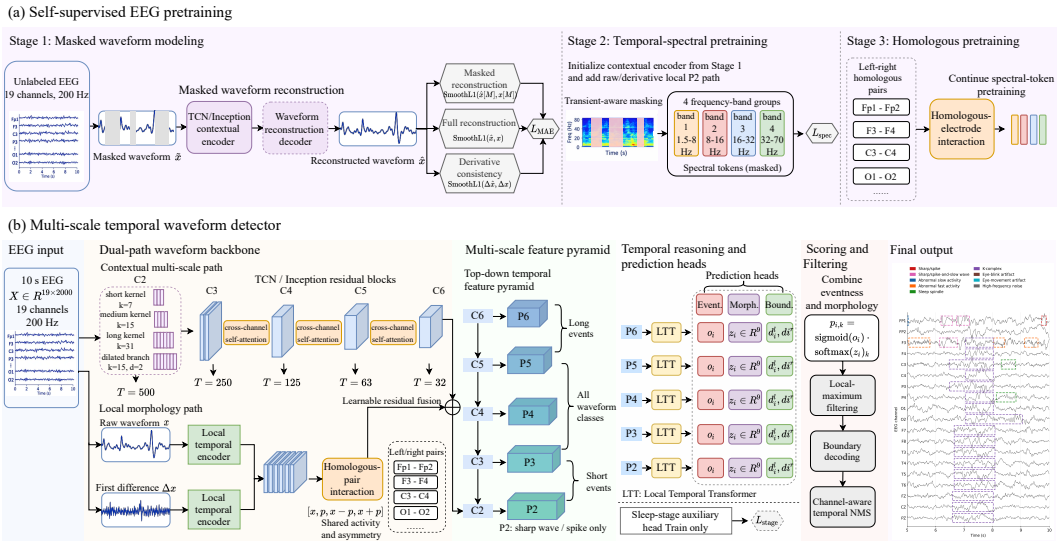


Figure 2: WA-Detector architecture and pretraining. (a) The backbone is pretrained through masked waveform reconstruction, masked spectral-token prediction, and homologous-electrode continuation. (b) A dual-path backbone combines contextual and raw/derivative morphology features with cross-channel interactions. A temporal feature pyramid and local Transformers feed heads for eventness, morphology, and boundary prediction. Predicted events are scored and filtered. The bottom-right panel shows an example of the detector output.

one of $K = 9$ waveform classes. WA-Detector outputs a set $\hat{\mathcal{E}}$ of scored tuples. We use temporal intersection-over-union (tIoU) to match predictions and annotations on the same channel. Figure 2 summarizes the self-supervised pretraining pipeline and the downstream detector.

4.2 Dual-Path Multi-Scale Backbone

The backbone combines a contextual path with a high-resolution morphology path. The contextual path is a TCN/Inception-style network (Ismail Fawaz et al., 2020). A strided temporal stem produces C2, and successive residual stages progressively reduce its temporal resolution to obtain C3–C6. Within each block, parallel branches capture short-, medium-, and long-range patterns together with a dilated temporal context.

The morphology path preserves short transients that may be attenuated by the contextual stem. It applies separate size-9, stride-2 convolutions to the raw waveform and its first temporal difference, concatenates the two feature maps, and uses a size-9, stride-2 separable temporal block to obtain a 500-step local feature L_2. This feature is merged with the contextual feature through a learnable residual gate, $\bar{\boldsymbol{C}}_2 = \boldsymbol{C}_2 + \tanh(\alpha) L_2$, where α is initialized to 0.1. The fused $\bar{\boldsymbol{C}}_2$ is the finest-resolution backbone feature supplied to the feature pyramid.

At C3–C5, four-head channel self-attention (Vaswani et al., 2017) treats the 19 electrodes as tokens independently at each temporal position and augments them with learned electrode embeddings. On the fused finest-resolution feature, a homologous-pair module transforms each valid left/right pair (x, p) using $[x, p, x-p, x+p]$ and adds the result through a second learnable residual gate. The module matches left-right pairs by electrode name and processes only complete pairs; other channels remain unchanged.

4.3 Feature Pyramid and Temporal Modeling

The backbone features are projected to a common representation and merged through a top-down feature pyramid (Lin et al., 2017a), yielding P2–P6. Because the finest level is intended for brief transients, P2 is supervised only for the merged sharp/spike class, while P3–P6 cover all nine waveform classes. After temporal positional encoding, a lightweight windowed Transformer models

temporal dependencies independently for each electrode at each pyramid level. The resulting features are passed to the prediction heads.

4.4 Decomposed Eventness and Training Objective

The point-based prediction head separates the presence of an event from its morphology. At each channel-specific pyramid location i, it predicts an eventness logit o_i, conditional morphology logits $\boldsymbol{z}_i \in \mathbb{R}^K$, and distances (d_i^l, d_i^r) to the left and right event boundaries. These distances define the event interval without predefined anchors (Zhang et al., 2022). The confidence assigned to class k is

$$p_{i,k} = \sigma(o_i)\,\mathrm{softmax}(\boldsymbol{z}_i)_k. \tag{1}$$

This factorization allows eventness to be learned from both positive and background locations, while morphology is classified only where an event is present. The training objective is

$$\mathcal{L} = \lambda_{\mathrm{evt}}\mathcal{L}_{\mathrm{evt}} + \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{morph}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{dist}} + \lambda_{\mathrm{IoU}}\mathcal{L}_{\mathrm{tIoU}} + \lambda_{\mathrm{stage}}\mathcal{L}_{\mathrm{stage}}. \tag{2}$$

Here $\mathcal{L}_{\mathrm{evt}}$ is focal loss (Lin et al., 2017b), $\mathcal{L}_{\mathrm{morph}}$ is cross-entropy over positive locations, and $\mathcal{L}_{\mathrm{dist}}$ and $\mathcal{L}_{\mathrm{tIoU}}$ supervise boundary distances and interval overlap. Their respective weights are 1, 2, 1, and 1. A location is positive only if it lies within an event, is near its center, and matches the duration range of its pyramid level. Finally, a training-only sleep-stage classifier on the coarsest pyramid feature provides the auxiliary cross-entropy $\mathcal{L}_{\mathrm{stage}}$ with weight 0.2.

4.5 Self-Supervised Pretraining and Fine-Tuning

Pretraining follows the incremental construction of the backbone. We first train the contextual path with masked waveform modeling. The decoder reconstructs masked samples, while derivative-consistency and full-waveform terms preserve transient shape and overall signal fidelity. We then initialize the dual-path backbone from this model, add the raw/derivative local path, and predict quantized spectral tokens at masked locations across four frequency bands. Temporal masks are preferentially placed around high-derivative or high-energy regions to emphasize transient morphology, and a small fraction of electrodes is additionally masked across the whole window so that a channel must be reconstructed from its neighbours. Finally, we add the homologous-electrode module and continue the spectral-token objective.

WA-Detector is fine-tuned end to end with Adam (Kingma & Ba, 2015) and class-balanced window sampling. Training augmentation includes channel dropout and permutation, temporal shifts, amplitude scaling, and additive noise. At inference, class scores are locally peak-filtered, decoded into temporal intervals, and consolidated using channel-aware non-maximum suppression. Detailed optimization and augmentation settings are provided in Appendix D. WA-Detector has approximately 5.14M inference parameters.

5 WaveAnchor-Agent

5.1 Evidence-Grounded Architecture

WA-Agent builds on LLM-based tool use (Yao et al., 2022), separating tool selection from response generation, as illustrated in Figure 3. A controlled registry resolves opaque study identifiers to EEG files and rejects arbitrary paths. The language-model backend is configurable; in this work, it is accessed through the DeepSeek API and invoked in two role-specific stages. Given the current request and dialogue history, the planner selects typed tools and supplies schema-validated arguments. The result is serialized as structured evidence, together with available provenance and limitations, for a separate responder call. Neither role directly accesses the EEG file or detector checkpoint; all signal-dependent claims must pass through the controlled tool interface.

5.2 Controlled Tool Set

WA-Agent exposes five typed tools. **Metadata inspection** returns recording duration, sampling rate, and available channels. **Background analysis** computes spectral peaks and amplitude summaries deterministically. **Waveform detection** uses WA-Detector (Section 4) to return event classes,

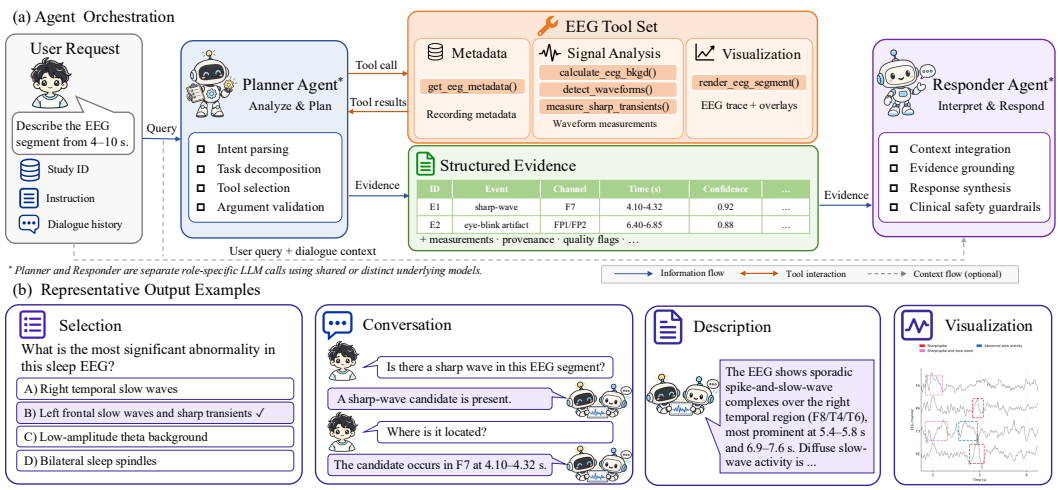


Figure 3: WA-Agent workflow and representative outputs. (a) The planner chooses among metadata, signal-analysis, and rendering tools. Structured results carry waveform classes, channels, time intervals, measurements, and plot references to a separate responder. Planner and responder may use the same or different language models. (b) Selection, conversation, description, and visualization examples come from different EEG segments.

channels, intervals, confidence scores, complete per-class summaries, and signal-quality findings. **Sharp-transient measurement** reports duration-based subtypes, baseline-to-peak and peak-to-peak amplitudes in microvolts, and measurement limitations for detected sharp waves and spike-and-slow-wave complexes. **Visualization** produces PNG plots of EEG segments with detected-event overlays.

5.3 Responder and Deterministic Guards

The responder receives the user request, dialogue history, planner note, and structured tool results. A compact finding table organizes detected events by class, channel, count, and time interval. The responder can also receive a rule-based text summary of detected waveforms, artifacts, and background measurements as additional evidence. We instruct the responder to describe EEG findings without making clinical diagnoses. After generation, rule-based checks preserve selection labels, verify channel counts and class–channel lookups, and add measurements when needed. Evidence configurations are detailed in Appendix F.

6 Experiments

6.1 WA-Detector

6.1.1 Detection Results

Table 1 reports WA-Detector's per-class AP at channel-matched tIoU 0.5, with a nine-class macro average (mAP@0.5) of 45.22%. High-frequency noise and lateral eye movement perform best, while abnormal slow and fast activity remain more challenging. Detection examples and a fixed-threshold confusion matrix are provided in Appendix E.

Figure 4 presents precision–recall curves for four waveform families. Artifacts have the highest family mAP (60.50%), followed by sharp-transient candidates (45.34%), sleep waveforms (38.73%), and abnormal activities (28.67%). Every class is recall-limited rather than precision-limited: precision stays above 0.8 over the highest-confidence detections, but the maximum recall remains well below one, ranging from 0.33 for abnormal fast activity to 0.84 for high-frequency noise. We also evaluate mAP across matching thresholds (Zhang et al., 2022), with family curves shown in the insets. Macro mAP decreases from 49.16% at tIoU 0.3 to 28.85% at 0.7, averaging 41.97% over tIoU 0.3–0.7.

Table 1: Per-class test AP at channel-matched tIoU 0.5 (%).

Class	AP@0.5
Sharp wave	35.07
Spike-and-slow wave	55.62
Abnormal slow wave	27.79
Abnormal fast activity	29.56
Sleep spindle	45.78
K-complex	31.68
Eye-blink artifact	52.11
Lateral eye movement	61.16
High-frequency noise	68.22
Macro mAP	**45.22**

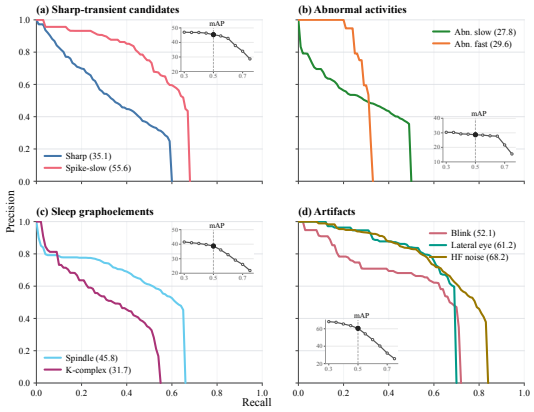


Figure 4: Precision–recall curves at tIoU 0.5. Parentheses: AP (%); insets: family mAP versus tIoU.

Table 2: Detector mAP@0.5 (%). †: published result.

Detector	WA-Bench	CerebraGloss-Bench
CerebraGloss-YOLO	34.56	40.95†
WA-Detector	**45.22**	**44.39**

Table 2 compares dataset-specific versions of WA-Detector and CerebraGloss-YOLO (Gu et al., 2026). On the WA-Bench test data, WA-Detector improves mAP by 10.66 percentage points, with gains in sharp-wave and spike-and-slow-wave detection, sleep waveforms, and abnormal fast activity. A separately trained version using the CerebraGloss-Bench label set also exceeds its published detector baseline. Class-wise results, label schemas, and evaluation settings are in Appendix I.

6.1.2 WA-Detector Ablations

We retrain variants using the same clinical splits, early-stopping rule, and inference settings, with matched pretraining checkpoints for backbone changes. Table 3 shows substantial mAP drops when pretraining, channel self-attention, or additional TU training data are removed. Per-class analysis indicates that the additional TU data primarily improve artifact detection, increasing artifact-family mAP from 30.51% to 60.50%. Full results are in Appendix E.

6.2 WA-Agent

The default configuration uses DeepSeek-V4-Flash for both language-model roles, three training-set demonstrations per role unless otherwise stated, and the same WA-Detector checkpoint with a reporting confidence threshold of 0.202. WA-Agent achieves complete required-tool coverage in the structured-evidence evaluation; detailed routing results are given in Appendix F.4.

6.2.1 Task Results

Scoring. Selection accuracy uses 175 four-option questions and synthetic silver answer keys. Description and conversation use case-specific EEG rubrics evaluated by DeepSeek-V4-Flash in an LLM-as-a-judge setting (Zheng et al., 2023). Strict scoring counts only pass judgments; partial-credit scoring assigns 1, 0.5, and 0 to pass, partial, and fail. Both scores are averaged over rubric items within each case and then equally over cases, with the two tasks reported separately. Re-scoring with DeepSeek-V4-Pro preserves the ordering of the main configurations for both tasks under strict and partial scoring. Scoring exceptions and full judge-sensitivity results are reported in Appendix F. Supplementary NLP scores and their limited alignment with evidence-grounded EEG interpretation are discussed in Appendix F.5.

Table 3: Detector ablations (%). Epi.: mean AP of sharp-wave and spike-and-slow-wave labels; Abn.: these labels plus abnormal slow and fast activity. Δ: mAP change from the full model (percentage points).

Variant	mAP@0.5	Epi. mAP	Abn. mAP	Δ
Full model	**45.22**	**45.35**	**37.01**	–
w/o pretraining	35.76	33.27	23.13	−9.46
w/o additional TU training data	34.59	35.44	34.77	−10.64
w/o channel self-attention	35.98	32.16	25.54	−9.24
w/o homologous-pair interaction	41.11	39.71	32.23	−4.11
w/o top-down pyramid fusion	41.41	35.31	29.29	−3.81
w/o local temporal attention	41.88	38.73	35.31	−3.34

Table 4: WA-Bench task results (%). Rubrics are strict/partial; Oracle uses annotated evidence.

System / evidence configuration	Selection	Description	Conversation
Direct image-based baselines			
GPT-5.5 (OpenAI, 2026)	**93.71**	21.99 / 31.59	19.58 / 33.46
Gemini 3.8 Flash (Google DeepMind, 2026)	84.57	26.09 / 36.23	19.22 / 30.40
Claude Sonnet 5 (Anthropic, 2026)	86.86	25.45 / 33.83	22.62 / 31.04
DeepSeek-V4.1-Flash (DeepSeek-AI, 2026)	90.29	20.21 / 28.13	12.54 / 20.58
CerebraGloss	61.71	18.31 / 25.84	6.25 / 17.36
WA-Agent: detected events			
Structured only	**93.71**	48.74 / 60.14	43.12 / 52.42
Summary priority, concise prompt	**93.71**	**59.00 / 68.73**	**48.10 / 55.52**
WA-Agent: Oracle evidence			
Structured only	97.71	71.63 / 79.75	77.58 / 83.10
Summary priority, concise prompt	98.29	91.88 / 94.84	93.66 / 94.86

Selection, description, and conversation. Table 4 includes off-the-shelf models and CerebraGloss (Gu et al., 2026) alongside WA-Agent on the 175 EEG cases in WA-Bench. The off-the-shelf models receive the benchmark task text and a plot of the corresponding EEG segment without waveform annotations or detection overlays, and generate answers directly from these inputs. Within each task, all systems are evaluated with the same factual rubric items and judge prompt. The detected-evidence rows use the WA-Detector checkpoint evaluated in Table 1, whereas CerebraGloss generates answers directly from EEG images. Structured detector evidence alone yields higher strict and partial-credit scores than all direct image-based baselines. In selection accuracy, GPT-5.5 matches this configuration (93.71%), while the remaining baselines score lower. Summary priority further improves description and conversation while preserving selection accuracy. Oracle conditions replace detected events with annotations and, when present, use annotation-derived summaries as an upper-bound control. These results support connecting localized waveform evidence and deterministic measurements to language generation. Additional cross-dataset results are provided in Appendix I. Example questions, agent responses, and reference answers are provided in Appendix G.1.

Visualization. The visualization task in WA-Bench covers 30 free-running two-turn conversations involving raw plots, detection overlays, and discussion. All 37 rendering requests successfully deliver valid, retrievable PNGs. Exact-call matching passes 58/60 turns and 28/30 conversations; the two exceptions are discussion turns with an additional background call.

6.2.2 WA-Agent Ablations

We examine detection, background analysis, and routing efficiency using the structured-evidence configuration (Table 5). Removing the detection interface reduces description and conversation scores under both strict and partial scoring, highlighting the importance of waveform-specific evidence for open-ended answers. Removing background analysis primarily affects description, where frequency and amplitude measurements complement the localized waveform findings. Fixed routing invokes detection, background analysis, and sharp-transient measurement for every request. The planner

Table 5: Structured-only WA-Agent ablations (%). D/C: description/conversation, strict/partial; Calls: calls/request.

Configuration	Selection	D rubric	C rubric	Calls
Structured-evidence baseline	**93.71**	**48.74 / 60.14**	43.12 / 52.42	1.637
w/o detection interface	86.29	29.18 / 34.65	14.53 / 17.05	**1.238**
w/o background tool	**93.71**	36.65 / 49.80	42.54 / 52.14	1.371
Fixed three-tool routing	93.14	46.94 / 59.38	**45.42 / 54.48**	3.000

Table 6: Role-specific model-size comparisons with detected evidence and summary priority (%). Rubrics are strict/partial.

Planner	Responder	Selection	Description	Conversation
Flash	Flash	**93.71**	**59.00 / 68.73**	48.10 / 55.52
Qwen3-8B	Flash	93.14	57.17 / 67.73	**50.59 / 58.01**
Flash	Qwen3-8B	78.29	45.33 / 59.65	42.34 / 54.58
Qwen3-32B	Qwen3-32B	87.43	51.51 / 63.80	45.84 / 56.07
Qwen3-8B	Qwen3-32B	86.86	52.15 / 64.32	44.65 / 55.30
Qwen3-32B	Qwen3-8B	74.86	45.57 / 60.20	41.52 / 53.78

reduces average calls from 3.000 to 1.637, a 45.43% reduction, with small changes in the description and conversation scores. Additional ablations are reported in Appendix F.

6.2.3 Planner and Responder Model Comparisons

We examine role-specific model-size reductions using Qwen3-8B and Qwen3-32B (Yang et al., 2025), with DeepSeek-V4-Flash (Flash) in both roles as the reference baseline. Table 6 reports Qwen pairings and single-role Flash/Qwen3-8B substitutions with detected evidence and concise summary-priority responses.

With a 32B responder, reducing the planner from 32B to 8B keeps scores close across all tasks. Keeping the 32B planner and reducing the responder to 8B produces larger drops, especially in selection accuracy. The Flash/Qwen3-8B substitutions show the same pattern: an 8B planner retains near-baseline answer quality, whereas an 8B responder lowers selection and description scores. These results support using a smaller planner while retaining a stronger responder. Across all pairings, structured-evidence runs achieve 99.19%–100.00% required-tool coverage and deliver all 37 requested PNGs in the visualization evaluation. The complete model matrix, execution settings, and detailed routing and rendering results are in Appendix H.

7 Conclusion

We introduced WA-Bench, a clinical EEG benchmark combining channel-level temporal annotations with case-specific atomic rubrics for waveform description and multi-turn dialogue. We also presented WA-Agent, a tool-using framework for fine-grained interpretation of multiple EEG waveform classes. WA-Detector localizes events by channel and time, and WA-Agent routes user requests to signal-analysis tools and explains the returned evidence. Experiments show effective waveform detection and reliable tool routing; selection accuracy and rubric-based evaluation on WA-Bench further support the quality of the generated answers. Several limitations remain. Our evaluation focuses on short, waveform-enriched segments and interictal interpretation; ictal analysis and routine continuous EEG are outside its scope. Waveform detection is also a bottleneck: missed or inaccurate detections can limit the completeness and accuracy of downstream answers. Finally, LLM-based rubric scoring may not fully capture clinical nuances or unsupported statements beyond the specified criteria. Future work will improve event recall and class discrimination, extend evaluation to continuous and ictal EEG, and complement rubric scoring with expert assessment of generated answers. Together, our benchmark and tool-using framework provide a foundation for transparent EEG assistants: generated explanations are linked to localized signal evidence, and their factual coverage can be assessed against case-specific criteria.

Ethics Statement

The EEG segments used in this study were selected from clinical patient recordings and annotated by an EEG specialist, with subsequent review by a medical trainee. All data were fully de-identified before use and contain no patient-identifying personal information. The system is a research prototype and is not intended for diagnosis or autonomous clinical decision-making. Model outputs may contain missed or false detections, and all displayed answers require review against the original recording by qualified professionals.

Reproducibility Statement

Sections 3–6 specify the task definitions, class schema, preprocessing, model modules, training objective, evaluation metrics, and result provenance. Detailed hyperparameters and implementation settings are provided in Appendix D. Upon acceptance, we will publicly release the code, pretrained model checkpoints, and a multi-class EEG waveform detection dataset comprising EEG segments, waveform annotations, and associated text data.

AI Use Statement

Generative AI tools were used to polish manuscript prose; support the retrieval and discovery of related work; assist research execution, primarily through coding and debugging; and generate initial drafts of the synthetic text datasets used in this study. Citations identified with AI assistance were checked against the original sources, and AI-assisted prose, code, and synthetic text were reviewed and revised by the authors before use. The authors take full responsibility for the resulting manuscript, code, data, and scientific claims.

References

Anthropic. System card: Claude Sonnet 5, 2026. URL https://www.anthropic.com/claude-sonnet-5-system-card.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/DeepSeek_V41_Tech_Report.pdf.

Google DeepMind. Gemini 3.8 Flash model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Wei Gu, Tianming Luo, Qiran Zhang, Mohan Ye, Xiao Shen, Wenxin Chen, Yunhuan Li, Yichen Zhang, Jing Hong, Bao-liang Lu, et al. CerebraGloss: Instruction-tuning a large vision-language model for fine-grained clinical EEG interpretation. In *International Conference on Learning Representations*, volume 2026, pp. 12594–12630, 2026.

John Guttag. CHB-MIT Scalp EEG Database. *PhysioNet*, June 2010. doi: 10.13026/C2K01R. URL https://doi.org/10.13026/C2K01R. Version 1.0.0.

Dengzhe Hou, Lingyu Jiang, Fangzhou Lin, and Kazunori D Yamada. CogEEGAgent: Toward autonomous cognitive EEG analysis with grounded execution and selection-aware verification. *arXiv preprint arXiv:2607.25045*, 2026.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *International Conference on Learning Representations*, volume 2025, pp. 55436–55457, 2025.

Jin Jing, Haoqi Sun, Jennifer A Kim, Aline Herlopian, Ioannis Karakis, Marcus Ng, Jonathan J Halford, Douglas Maus, Fonda Chan, Marjan Dolatshahi, et al. Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation. *JAMA Neurology*, 77(1):103–108, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Mustafa Aykut Kural, Lene Duez, Vibeke Sejer Hansen, Pål G Larsson, Stefan Rampp, Reinhard Schulz, Hatice Tankisi, Richard Wennberg, Bo M Bibby, Michael Scherg, et al. Criteria for defining interictal epileptiform discharges in EEG: A clinical validation study. *Neurology*, 94(20): e2139–e2147, 2020.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944. IEEE, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017b.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Silvia Lopez, G Suarez, D Jungreis, I Obeid, and Joseph Picone. Automated identification of abnormal adult EEGs. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5. IEEE, 2015.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

OpenAI. GPT-5.5 system card, 2026. URL `https://deploymentsafety.openai.com/gpt-5-5`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

Vinit Shah, Eva von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The Temple University Hospital seizure detection corpus. *Frontiers in Neuroinformatics*, 12:83, 2018.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9626–9635. IEEE, 2019.

Marleen C Tjepkema-Cloostermans, Rafael CV de Carvalho, and Michel JAM van Putten. Deep learning for detection of focal epileptiform discharges from scalp EEG recordings. *Clinical Neurophysiology*, 129(10):2191–2196, 2018.

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

Lillian Veloso, J McHugh, Eva von Weltin, Silvia Lopez, Iyad Obeid, and Joseph Picone. Big data resources for EEGs: Enabling deep learning research. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–3. IEEE, 2017.

Eva von Weltin, Tameem Ahsan, Vinit Shah, Dawer Jamshed, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. Electroencephalographic slowing: A primary source of error in automatic seizure detection. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5. IEEE, 2017.

Guoan Wang, Shihao Yang, and Feng Liu. NeuroWeaver: An autonomous evolutionary agent for exploring the programmatic space of EEG analysis pipelines. *arXiv preprint arXiv:2602.13473*, 2026.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, volume 2025, pp. 75310–75346, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *NeurIPS Workshop on Foundation Models for Decision Making*, 2022.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pp. 492–510. Springer, 2022.

Sha Zhao, Mingyi Peng, Haiteng Jiang, Tao Li, and Shijian Li. EEGAgent: A unified framework for automated EEG analysis using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18063–18071, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Shijian Li, and Gang Pan. BrainAgent: A large language model-driven multi-agent framework for autonomous brain signal understanding. *arXiv preprint arXiv:2606.25400*, 2026.

A Dataset and WA-Bench Details

This appendix provides additional details on data processing and WA-Bench.

A.1 Clinical Cohort and Split Statistics

The clinical data were acquired at a single hospital at 200 Hz using the international 10–20 system. In total, the dataset contains 34 training EDF recordings and 26 test EDF recordings. The training and test sets are separated at the patient level. Table 7 summarizes the processed data.

Table 7: Summary of the supervised clinical EEG data. "Distinct test time" is computed after merging overlapping windows within each source recording.

Statistic	Training	Test
Source recordings	34	26
Processed arrays/windows	34	175
Nominal signal duration (s)	8,877	1,750
Distinct test time (s)	–	1,648.5
Raw channel-level annotations	31,255	8,926
Nine-class annotations	28,634	7,344

The 34 processed training recordings contain 8,877 s (2.47 h) of signal. Their durations range from 14 to 2,398 s, with a median of 92.5 s. The test set contains 175 fixed 10-s windows. Because a residual interval at the end of an annotated epoch is covered by an end-aligned window, some adjacent test windows overlap. The nominal test duration is therefore 1,750 s, while the union of source-level time intervals is 1,648.5 s.

A.2 Preprocessing

We retain the following 19 electrodes from the international 10–20 system: FP1, FP2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, FZ, CZ, and PZ. Each EDF is band-pass filtered from 1.6 to 70 Hz, notch filtered at 50 Hz, resampled to 200 Hz, converted to an average reference, and expressed in microvolts.

A.3 Annotation and Sleep Stage

An EEG specialist annotated the sleep stage at 60-s intervals, and a medical trainee with EEG expertise subsequently reviewed the annotations. Each extracted 10-s test segment falls entirely within one annotated stage interval; no test window crosses a stage boundary. Table 8 shows the sleep-stage distribution of the test set.

The source training annotations contain 54 wake, 17 NREM I, 46 NREM II, 9 NREM III, and 1 REM annotations. Each count corresponds to a 60-s sleep-stage annotation and is therefore not directly comparable with the number of 10-s test windows in Table 8.

Table 8: Sleep-stage distribution of the 175 clinical test windows, reconstructed from the source annotations and absolute window boundaries.

Stage	Test windows
Wake	45
NREM I	47
NREM II	73
NREM III	10
Total	175

A.4 Label Normalization and Class Distribution

The fixed nine-class taxonomy contains sharp wave, spike-and-slow-wave complex, abnormal slow wave, abnormal fast activity, sleep spindle, K-complex, eye-blink artifact, lateral eye-movement artifact, and high-frequency noise. Table 9 gives the resulting channel-level event counts.

The distribution is long-tailed. Abnormal slow waves constitute 44.7% and 40.5% of the training and test annotations, respectively, whereas abnormal fast activity constitutes only 3.0% and 1.0%. The

Table 9: Distribution of channel-level temporal annotations under the nine-class taxonomy. Percentages are computed within each split.

Waveform class	Training		Test	
	Count	%	Count	%
Abnormal slow wave	12,805	44.72	2,977	40.54
Sharp wave	4,141	14.46	1,071	14.58
Sleep spindle	3,331	11.63	1,476	20.10
High-frequency noise	2,826	9.87	714	9.72
Eye-blink artifact	2,422	8.46	217	2.95
K-complex	977	3.41	261	3.55
Abnormal fast activity	849	2.97	73	0.99
Lateral eye-movement artifact	681	2.38	222	3.02
Spike-and-slow-wave complex	602	2.10	333	4.53
Total	28,634	100.0	7,344	100.0

dataset intentionally samples waveform-enriched regions; consequently, these class proportions do not reflect their prevalence in routine clinical EEG.

A.5 Supplementary Public Training Data

To increase waveform diversity during supervised learning, we additionally include a comparatively small set of training examples from TUAB (Lopez et al., 2015), TUEV (Obeid & Picone, 2016), TUSL (von Weltin et al., 2017), and TUEP (Veloso et al., 2017), collectively referred to as the TU dataset. After filtering and annotation, the TU dataset contains 1,871 samples and 31,496 channel-level annotations under the same nine-class taxonomy. Table 10 reports the resulting distribution.

Table 10: Distribution of the 31,496 channel-level annotations in the supplementary TU training dataset after normalization to the nine-class taxonomy.

Waveform class	Count	%
Sharp wave	8,879	28.19
Spike-and-slow-wave complex	933	2.96
Sleep spindle	2,708	8.60
K-complex	230	0.73
Eye-blink artifact	4,532	14.39
High-frequency noise	7,235	22.97
Abnormal slow wave	5,428	17.23
Abnormal fast activity	484	1.54
Lateral eye-movement artifact	1,067	3.39
Total	31,496	100.0

The raw annotation counts do not directly determine the contribution of each dataset during training. The WA dataset is segmented with a 1-s stride during training, allowing annotations to be reused across multiple overlapping training windows, whereas the TU dataset is segmented into non-overlapping windows. Consequently, TU annotations account for approximately 25.91% of the label instances encountered during joint training. The TU data are used only for training: the held-out detection set, all language tasks, and the waveform-visualization task are constructed exclusively from the clinical cohort described above.

A.6 WA-Bench Agent Tasks

The selection, description, and conversation tasks each cover the same 175 held-out windows without missing or duplicated cases. The waveform-visualization task uses a 30-case subset of these windows. For the language tasks, initial task text and, where applicable, atomic rubric criteria were generated from structured waveform evidence with AI assistance. A medical trainee with EEG expertise

manually checked all benchmark questions, selection options and answer keys, reference answers, and rubric criteria. The four tasks probe related but distinct abilities:

- **Selection** presents four options and requires the agent to identify the single correct answer.
- **Description** requires a concise synthesis of waveform type, location, timing, and artifacts.
- **Conversation** asks progressively more specific questions about the same segment and requires precise extraction of the waveform features relevant to each turn.
- **Waveform visualization** evaluates whether the agent selects and sequences the appropriate waveform-rendering and analysis tools.

Table 11 summarizes the benchmark. A conversation "turn" denotes one user–agent exchange. Of the 175 conversation cases, 160 contain three turns and 15 contain two turns, for 510 turns in total; one turn has no rubric criteria, leaving 509 rubric-scored turns. The description rubric contains 801 atomic criteria, or 2–9 per case (median 4). The conversation rubric contains 749 criteria, or 1–8 per case (median 4), distributed across the 509 rubric-scored turns. Atomic criteria separately assess facts such as waveform identity, laterality, channel or region, temporal range, and artifact context.

Table 11: Composition of the segment-level agent benchmark.

Task	Cases	Turns	Atomic criteria
Selection	175	175	–
Description	175	175	801
Conversation	175	510	749
Waveform visualization	30	60	–

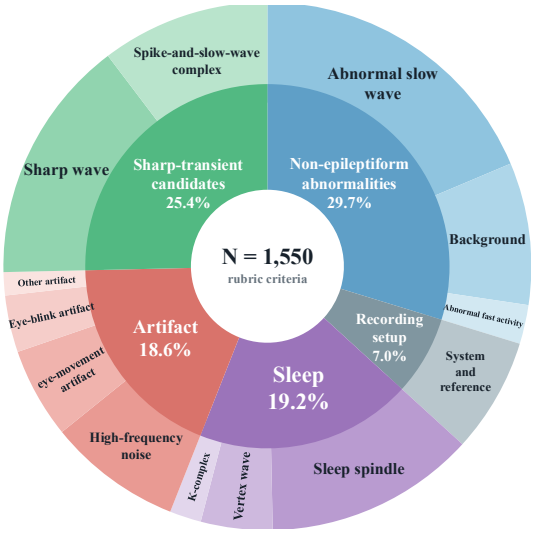


Figure 5: Detailed semantic distribution of the 1,550 atomic rubric criteria in WA-Bench's description and conversation tasks. The inner ring shows five major topics, while the outer ring resolves them into 13 detailed EEG concepts. Sector area is proportional to criterion count.

The main-text overview summarizes the five major semantic groups represented in the description and conversation rubrics. Figure 5 provides the corresponding fine-grained breakdown. For this descriptive analysis, each atomic criterion is assigned to one primary topic according to the waveform or EEG property that it explicitly evaluates. The inner ring aggregates all 1,550 criteria into non-epileptiform abnormalities (29.7%), sharp-transient candidates (25.4%), sleep (19.2%), artifact (18.6%), and recording setup (7.0%). The outer ring resolves these groups into 13 detailed concepts, including abnormal slow and fast activity, background activity, epileptiform transients, physiological sleep graphoelements, common artifacts, and the recording montage and reference. Sector area is proportional to the number of atomic criteria.

Non-epileptiform abnormalities and sharp-transient candidates together account for 55.1% of the criteria, reflecting the benchmark's emphasis on recognizing clinically salient waveform findings. Sleep and artifact criteria remain substantial, requiring the agent to distinguish physiological sleep waveforms and common artifacts from abnormal activity. Recording-setup criteria form the smallest group and test whether the response correctly identifies the channel system and reference configuration. The benchmark therefore assesses a broad range of segment-level EEG findings rather than only epileptiform events.

A.7 Rubric Scoring

Description and conversation are evaluated using an API-based judge, with each criterion assessed independently. The judge evaluates the agent response against the corresponding rubric and assigns a score to every atomic criterion.

For each criterion, the judge assigns one of three outcomes. `Pass` indicates that every material fact is supported by the response. `Partial` indicates that the primary finding—the waveform type and its laterality or anatomical region—is correct, but a secondary detail, such as one of several channels, a time interval, frequency, or amplitude, is incomplete or imprecise. `Fail` indicates that the primary finding is absent, contradicted, or incorrect. Semantic equivalence is accepted: electrode names and anatomical-region descriptions are treated as equivalent when they denote the same locations in the international 10–20 system, and modest rounding of temporal boundaries is allowed when the intervals substantially agree. Style, diagnostic conclusions, recommendations, and safety disclaimers do not affect the judgment.

For the primary strict score, `Pass` receives one point, while `Partial` and `Fail` receive zero. We additionally report a partial-credit score in which `Partial` receives half a point. Criterion scores are first averaged within each case. Task performance is then computed by averaging the case scores, giving every case equal weight. The overall semantic score is the equal-weight average of the description and conversation task scores.

A.8 Separation of Detection, Semantic, and Routing Evaluation

The evaluation suite distinguishes three notions of correctness. Detector mAP compares predicted channel-level temporal intervals with physician annotations and measures waveform classification and localization. The description and conversation rubrics instead measure whether the agent extracts and communicates the clinically relevant facts for a complete 10-s segment. An answer may correctly describe the main finding without listing every annotated event, whereas a detector must localize individual events to score well. Routing evaluation measures whether the agent selects the appropriate waveform-rendering or analysis tool and invokes the required tools in the correct sequence; it does not assess the localization accuracy of boxes displayed in the resulting visualization. These evaluation layers are reported separately throughout the experiments: mAP for fine-grained temporal detection, rubric-based scores for segment-level interpretation, and routing accuracy for tool-use behavior.

B Rationale for WA-Detector's Nine-Class Waveform Taxonomy

B.1 Design Objective

The nine classes cover four complementary sources of evidence:

1. **Sharp-transient candidates:** the sharp-wave class includes sharp waves, spikes, and vertex sharp transients, while spike-and-slow-wave complexes form a separate class. Vertex sharp transients are benign physiological waveforms of sleep, but their local morphology can be nearly identical to that of sharp waves. In clinical interpretation, the distinction is often made from contextual evidence—including sleep stage, spatial distribution, bilateral symmetry, and accompanying sleep graphoelements—rather than from the isolated transient alone. We therefore do not require the waveform detector to separate vertex transients from sharp waves. This choice is also consistent with CerebraGloss, which assigns vertex waves to its sharp-wave category.
2. **Non-epileptiform abnormalities:** abnormal slow waves and abnormal fast activity represent focal or regional deviations that are important for waveform description but are not covered by an epileptiform-only ontology.
3. **Physiological sleep graphoelements:** sleep spindles and K-complexes provide explicit normal sleep targets. Detecting them is also important for avoiding confusion between physiological transients and abnormal discharges.
4. **Common confounders:** eye-blink artifact, lateral eye-movement artifact, and high-frequency noise represent frequent non-cerebral events that can mimic or obscure pathologi-

Table 12: Principal differences between the WA-Detector nine-class taxonomy and CerebraGloss.

Dimension	WA-Detector taxonomy	CerebraGloss
Sharp wave / spike / vertex transient	One detector class; the Agent subsequently evaluates sleep and spatial context to identify possible vertex transients	Separate sharp-wave and spike categories; vertex waves are included in the sharp-wave category
Eye movement	Transient potential change at the moment of eye movement	Full leftward or rightward eye-movement interval
Slow / fast activity	Abnormal slow waves and abnormal fast activity are explicit targets	Neither is an independent target
Artifact-induced sharp transient	Excluded from the sharp-wave category	May be included as a sharp wave, including blink-related deflections

cal waveforms. Modeling these classes allows the detector and downstream agent to qualify evidence rather than treating every salient transient as cerebral activity.

This composition gives the downstream agent access to abnormal findings, normal sleep context, and common technical confounders within a single temporal detection interface. Categories that are outside this scope are excluded. The treatment of vertex transients reflects a deliberately staged design. In the first stage, WA-Detector groups morphologically similar sharp transients into a common sharp-wave class. In the second stage, the Agent is explicitly prompted to reconsider central or midline sharp-wave detections in light of sleep stage, spatial distribution, symmetry, and co-occurring sleep spindles or K-complexes, and to describe them as possible vertex sharp transients when supported by that context (Appendix J.2). Thus, the detector preserves sensitivity to the shared morphology, while the Agent performs the contextual distinction needed for clinical interpretation.

B.2 Difference from the CerebraGloss Label Space

CerebraGloss (Gu et al., 2026) also defines nine waveform categories. It shares our convention of placing vertex waves in the sharp-wave category, but its other annotation targets and temporal boundaries differ from ours in four substantive ways:

1. **Sharp waves, spikes, and vertex transients.** We use one detector class for sharp waves, spikes, and vertex sharp transients, and defer the distinction between a possible vertex transient and a pathologic sharp wave to the Agent's contextual reasoning stage. CerebraGloss instead retains separate sharp-wave and spike categories, although it likewise includes vertex waves in the sharp-wave category.
2. **Eye-movement artifacts.** CerebraGloss marks the complete duration of a leftward or rightward eye movement. Such intervals can range from 1–2 s to 5–6 s or longer, producing targets with highly variable duration and morphology. We instead annotate the transient potential change at the moment of eye movement. This produces more consistent temporal boundaries and a more uniform target morphology.
3. **Abnormal slow and fast activity.** We add abnormal slow waves and abnormal fast activity as explicit targets. Although neither category alone establishes epileptic activity, both represent clinically important findings that should be retained in waveform descriptions.
4. **Artifact-induced sharp transients.** CerebraGloss may label sharp transients caused by artifacts, including blink-related sharp deflections, as sharp waves. We exclude artifact-induced sharp transients from the sharp-wave category because their meaning differs fundamentally from genuine cerebral sharp waves and many can be distinguished morphologically.

Table 12 summarizes these design differences. Because these differences affect both class semantics and temporal boundaries, the two taxonomies are not interchangeable at the level of output indices or event intervals. Comparisons must either map predictions to an explicitly declared common ontology or evaluate each model under its native annotation protocol.

Figure 6 is intended to show how the two datasets differ in their annotation of blink artifacts and the accompanying sharp-like deflections. The left example is drawn from CerebraGloss-Bench. Its

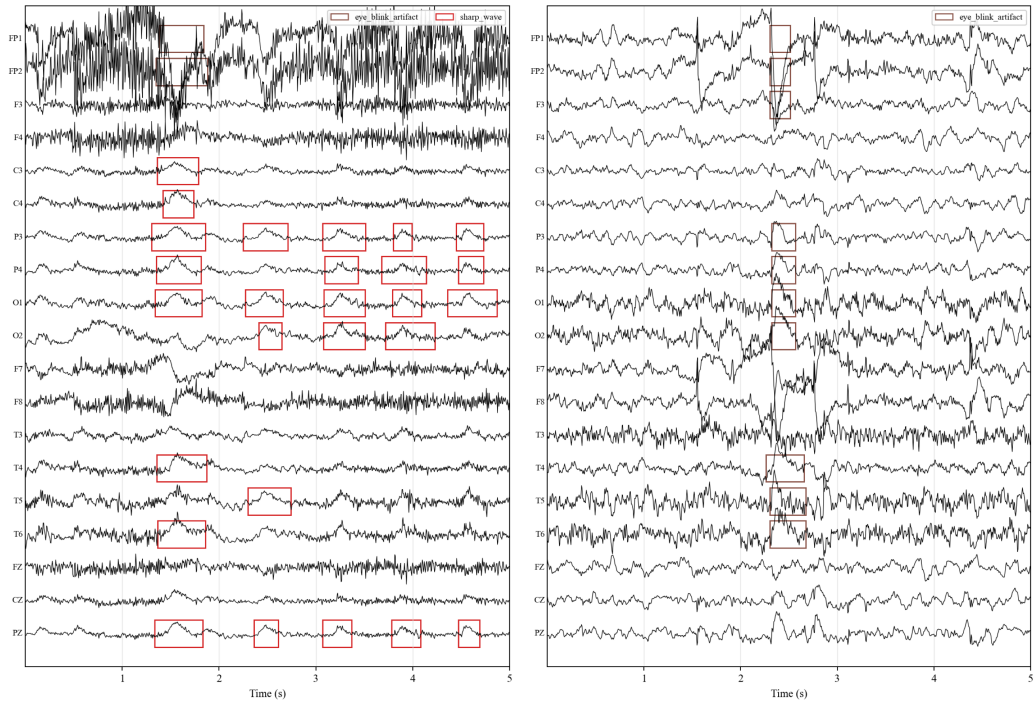


Figure 6: Blink-artifact and associated sharp-wave annotation differences. The left excerpt is from CerebraGloss-Bench, where the blink is labeled as `eyem` on FP1/FP2 but synchronous deflections over other channels are labeled as sharp waves. The right excerpt is from our test set, where the propagated event is labeled consistently as eye-blink artifact across the affected channels. Brown boxes denote eye-blink artifact and red boxes denote sharp-wave annotations. The comparison is intended to show both the difference in blink-artifact coverage and its consequence for sharp-wave supervision: the red-boxed events on the left are time-locked to the blink and are not visually typical of isolated cerebral sharp waves. Both panels show 5-s excerpts with identical temporal scale.

frontopolar blink is annotated as `eyem` on FP1 and FP2, whereas the time-locked deflections on eight non-frontopolar channels are assigned the sharp-wave label. These deflections are visually distinct from isolated cerebral sharp waves. The right example is drawn from our test set. Here, the propagated event is consistently represented as a blink artifact on 10 of the 19 channels rather than producing additional sharp-wave targets.

B.3 Scope and Limitations

The taxonomy intentionally prioritizes reproducible morphology at the detection stage rather than forcing the detector to make distinctions that depend on broader clinical context. A detector label identifies a visual waveform candidate rather than assigning a definitive cause. For example, abnormal slow activity is localized as a waveform, but the label does not assert a structural lesion. Likewise, the sharp-wave class contains sharp waves, spikes, and benign sleep-related vertex sharp transients; the detector output alone therefore establishes neither epileptiformity nor a diagnosis of epilepsy or an ongoing seizure. Artifact labels similarly describe signal contamination without identifying its definitive physical cause.

This limitation is handled explicitly by the Agent in the second stage, rather than through an additional detector class. The Responder Prompt instructs the Agent that the detector's sharp-wave class is heterogeneous and that central or midline sharp waves, particularly during sleep or alongside sleep graphoelements, may represent vertex sharp transients. The Agent can therefore use sleep staging, spatial distribution, bilateral symmetry, and co-occurring graphoelements to separate likely vertex transients in its description. These contextual cues remain interpretive evidence rather than a

deterministic relabeling rule: no single feature, such as location or sleep state alone, can definitively identify every sharp transient.

The nine classes are also not intended to cover every pattern present in routine EEG. Rhythmic and periodic patterns, additional benign variants, electrode-specific failures, and background-state descriptors are not independent detector targets. Expanding the ontology would require additional annotation guidelines and sufficient examples per class; otherwise, a larger label space could increase ambiguity without improving downstream evidence quality.

C Rubric Construction, Scoring, and Worked Example

C.1 Overview

The rubrics in WA-Bench are constructed from held-out 10-s clinical EEG segments through an AI-assisted, human-reviewed pipeline. Their purpose is to measure, in the conversation and description tasks, whether an agent states the waveform facts specified in the reference answer, rather than whether it reproduces the reference wording. Figure 7 summarizes the construction process.

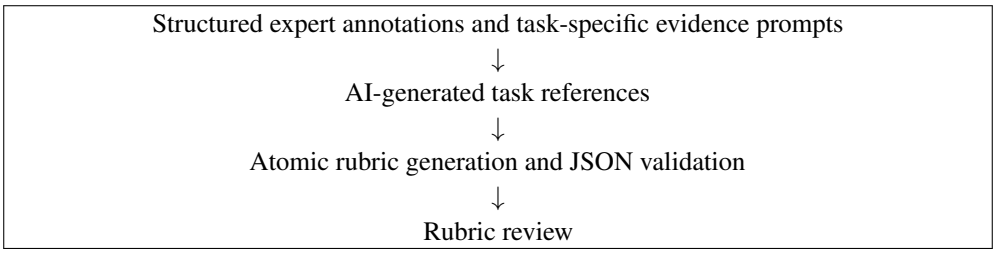


Figure 7: AI-assisted construction of WA-Bench rubrics. A medical trainee with EEG expertise reviews all rubric criteria.

The construction procedure is as follows:

1. **Evidence preparation.** Each case is linked to one held-out waveform window and its structured evidence, including waveform class, channel, interval, sleep context when available, and auxiliary signal descriptions.
2. **Reference generation.** Task-specific prompts produce a multiple-choice item, a single-paragraph description, and a progressive multi-turn conversation.
3. **Rubric generation.** The reviewed description or the current turn of a reviewed conversation is supplied to a rubric-generation model. The model decomposes the reference into a minimal set of independently judgeable waveform facts. Outputs are parsed as JSON.
4. **Rubric review.** The same trainee manually checks every generated criterion before finalization. Reviewed criteria are stored separately from evaluation inputs. The final benchmark contains 801 description criteria and 749 conversation criteria.

C.2 Rubric Semantics

Each criterion represents one factual waveform requirement that can be evaluated independently. Depending on the user request and reference response, a criterion may cover waveform identity, frequency, amplitude, temporal range, channel or regional localization, laterality, background activity, normal sleep graphoelements, or artifacts. Closely coupled properties may be grouped into one criterion, while independently answerable facts are separated.

The rubric deliberately excludes diagnostic conclusions, epileptogenic-focus claims, clinical significance, prognosis, treatment, and recommendations. This restriction separates factual evidence extraction from higher-level clinical inference.

C.3 Canonical Data Formats

Each description rubric is stored as one JSONL record:

Description Rubric Record


```
{
  "task_id": "<case id>",
  "filename": "<10-s waveform file>",
  "prompt_type": "description",
  "rubric": {
    "turn1": [
      {
        "id": "Q1",
        "criterion": "<independently judgeable waveform fact>",
        "reference_status": "pass"
      }
    ]
  },
  "status": "success"
}
```


Conversation rubrics use the same criterion schema but retain separate lists for each turn:

Conversation Rubric Record


```
{
  "task_id": "<case id>",
  "filename": "<10-s waveform file>",
  "prompt_type": "conversation",
  "rubric": {
    "turn1": [{"id": "Q1", "criterion": "...", "reference_status": "pass"}],
    "turn2": [{"id": "Q1", "criterion": "...", "reference_status": "pass"}]
  },
  "status": "success"
}
```


Rubric-generation input templates and system prompt excerpts are provided in Appendix J.5.

C.4 Worked Rubric Example

This example traces one WA-Bench case through its reference answer and reviewed atomic rubric. The case corresponds to a 10-s segment containing right-temporal spike-and-slow-wave complexes, widespread slow activity, and limited ocular and high-frequency artifacts.

C.4.1 Conversation Task

The conversation task is formulated as a progressive three-turn dialogue.

Turn 1

User: What are the main waveforms in this 10-second segment, and how are they distributed?
Reference agent: Two main waveforms are present. Spike-and-slow-wave complexes are restricted to the right temporal region: F8 at 6.88--7.54 s, T4 at 7.02--7.55 s, and T6 at 5.38--5.83 s and again at 7.06--7.56 s. Slow waves are widespread, with the greatest concentration over the right temporal region and additional involvement of the left temporal, bilateral frontal, right parietal, central, and occipital regions.

The turn-1 rubric is:

- **Q1:** The response identifies spike-and-slow-wave complexes restricted to the right temporal region.
- **Q2:** The response identifies widespread slow waves, maximal over the right temporal region, with additional left temporal, bilateral frontal, right parietal, central, and occipital involvement.

Turn 2

User: Can you detail the distribution and laterality of the spike-and-slow-wave complexes?
Reference agent: The spike-and-slow-wave complexes are confined to the right temporal region, involving F8, T4, and T6, with no involvement of the left hemisphere. T6 shows the earliest occurrence at 5.38--5.83 s. A later cluster involves all three electrodes between 6.88 and 7.56 s.

The turn-2 rubric contains one criterion:

- **Q1:** The response identifies spike-and-slow-wave complexes confined to the right temporal region.

Although the reference provides electrode-level timing, the user primarily asks about distribution and laterality. The final rubric therefore retains the central right-temporal localization requirement and does not require every timestamp to be reproduced.

Turn 3

User: Where is the high-frequency noise, and are there any eye-movement artifacts?
Reference agent: High-frequency noise is seen over the right occipital region (O2) during 0.03--1.19 s and over the left temporal region (T3) during 0.92--6.65 s. Eye-movement artifact is present at the frontopolar leads FP1 and FP2 during 0.22--0.49 s and 6.08--6.31 s. An additional artifact is recorded over T4 during 2.53--3.24 s.

The turn-3 rubric is:

- **Q1:** The response identifies high-frequency noise over the right occipital and left temporal regions.
- **Q2:** The response identifies eye-movement artifact at the frontopolar leads FP1 and FP2.

The final conversation rubric for the example is serialized as follows:

Serialized Conversation Rubric


```
{
  "task_id": "0",
  "prompt_type": "conversation",
  "rubric": {
    "turn1": [
      {"id": "Q1", "criterion": "The description identifies
      spike-and-slow-wave complexes restricted to the right temporal region.",
      "reference_status": "pass"},
      {"id": "Q2", "criterion": "The description identifies widespread slow
      waves, maximally over the right temporal region, also involving left
      temporal, bilateral frontal, right parietal, central, and occipital
      regions.", "reference_status": "pass"}
    ],
    "turn2": [
      {"id": "Q1", "criterion": "The description identifies
      spike-and-slow-wave complexes confined to the right temporal region.",
      "reference_status": "pass"}
    ],
    "turn3": [
      {"id": "Q1", "criterion": "The description identifies high-frequency
      noise over the right occipital and left temporal regions.",
      "reference_status": "pass"},
      {"id": "Q2", "criterion": "The description identifies eye-movement
      artifact at the frontopolar leads FP1 and FP2.", "reference_status":
      "pass"}
    ]
  },
  "status": "success"
}
```

C.4.2 DESCRIPTION TASK

The reference description is:

Reference EEG Description

```
The channel layout follows the standard 10--20 system with an average
reference.  Overall data quality is good, with high-frequency noise over
the right occipital region (O2) during 0.03--1.19 s and over the left
mid-temporal region (T3) during 0.92--6.65 s.  Frontopolar blink artifact
is present at 0.22--0.49 s and 6.08--6.31 s.  Slow waves are frequent and
most prominent over the right temporal region, with additional involvement of
the left temporal, bilateral frontal, central, and occipital regions, and the
right parietal region, occurring at multiple intervals throughout the segment.
Spike-and-slow-wave complexes are observed over the right temporal region at
approximately 5.38--5.83 s and 6.88--7.56 s.
```

The reviewed description rubric contains five criteria:

Table 13: Atomic description criteria for the worked example.

ID	Criterion
Q1	Channel layout follows the standard 10–20 system with average reference.
Q2	High-frequency noise is identified over the right occipital region at approximately 0.0–1.1 s and the left mid-temporal region at approximately 0.9–6.6 s.
Q3	Frontopolar blink artifact is identified at approximately 0.2–0.4 s and 6.0–6.3 s.
Q4	Slow waves are identified as most prominent over the right temporal region, with additional involvement of the left temporal, bilateral frontal, central, occipital, and right parietal regions.
Q5	Spike-and-slow-wave complexes are identified over the right temporal region.

This decomposition illustrates the intended granularity. The two artifact observations are independently judgeable and therefore receive separate criteria. The slow-wave criterion groups its principal and secondary spatial distributions because they jointly describe one widespread finding. Exact occurrence times for every slow-wave box are not converted into separate criteria. The rubric also omits the diagnostic implications of the right-temporal discharge because the test set is intended to assess factual waveform description rather than clinical diagnosis.

D IMPLEMENTATION DETAILS

D.1 WA-DETECTOR CONFIGURATION

Table 14: WA-Detector architecture and optimization settings.

Component	Setting
Input	19 channels, 200 Hz, 10 s
Backbone	Dual-path TCN/Inception, base width 48
Morphology path	Raw waveform + first difference
Cross-channel modules	Channel self-attention + homologous pairs
Pyramid	P2–P6, 128 channels
Temporal head	1 local block, window 64, 4 heads
Output	Eventness + 9-class morphology + boundaries
Pretraining	Masked waveform, spectral tokens, homologous continuation
Optimizer	Adam
Learning rate	1.5×10^{-4}
Effective batch size	64
Batch size / accumulation steps	16 / 4
Weight decay	0
Learning-rate decay factor / minimum	0.991 / 5×10^{-6}
Random seed	1

D.2 Training Augmentation

The five augmentations below are applied to a training window in the order listed, each independently with probability 0.5. None is applied during testing or inference.

- **Channel dropout.** Between one and six randomly chosen electrodes are zeroed and marked as missing, subject to retaining at least three non-missing electrodes. Dropped electrodes are removed from each annotation, and annotations with no remaining electrode are discarded.
- **Channel permutation.** The 19 rows are randomly permuted, and the missing-channel mask and the electrode position indices are permuted with them, so electrode identity is preserved while the input order changes.
- **Temporal shift.** The window is circularly rolled by an offset drawn uniformly from $[-20\%, 20\%]$ of the window length, that is $[-2, 2]$ s, and annotation boundaries are shifted with it.
- **Amplitude scaling.** The window is multiplied by a factor drawn uniformly from $[0.9, 1.1]$.
- **Additive noise.** Gaussian noise with standard deviation 0.02 is added to the per-channel standardized signal.

E Additional WA-Detector Results and Analysis

E.1 Ablation Studies

E.1.1 Ablation Settings

We describe the additional settings and initialization controls used in the ablation study below.

The optional P1 level. P1 is an additional pyramid level below P2. It carries 1000 temporal positions for a 10-second window, one every 10 ms, compared with 500 positions, or one every 20 ms, at P2. It uses its own lightweight prediction head with 64 hidden channels and, like P2, is supervised only for the merged sharp/spike class. Events shorter than 0.1 s are assigned to it, so its duration range deliberately overlaps that of P2, which covers events up to 0.15 s. Enabling P1 requires P2 and roughly doubles the total number of pyramid positions, from 970 to 1970; the corresponding run therefore uses a batch size of 8 with 8 gradient-accumulation steps, keeping the effective batch size at 64. The full model leaves P1 disabled.

Channel masking during pretraining. In addition to the temporal masks of Section 4, each electrode is independently selected with probability 0.1 and masked across the entire window. Reconstructing a fully masked electrode requires information from the remaining ones, which discourages the backbone from treating channels independently. The two corresponding ablations set this probability to 0 and to 0.2.

The two homologous-pair initialization controls. Both variants leave the homologous-pair module randomly initialized when fine-tuning starts, and they differ in how much pretraining the rest of the backbone has received. The direct variant starts from the dual-path backbone before the homologous-electrode pretraining stage and skips that stage. The budget-matched variant instead repeats the same 40-epoch spectral-token continuation with the homologous-pair module absent, so the backbone receives the same amount of pretraining and only the pretrained homologous-pair weights are withheld. Comparing the two separates the contribution of those weights, 1.86 points of mAP, from the contribution of the additional pretraining epochs, 1.36 points.

E.1.2 Complete Ablation Results

Tables 15 and 16 report every completed detector ablation. The experiments use the same training and test splits and test mAP@0.5 for early stopping. Architecture variants that change the available backbone modules use correspondingly modified pretraining checkpoints. Epi. mAP averages the sharp-wave class and spike-and-slow-wave complexes; Abn. mAP averages those two together

Table 15: Architecture and pretraining ablations (%). Epi.: mean AP of sharp-wave and spike-and-slow-wave labels.

Variant	mAP@0.5	Epi. mAP	Abn. mAP	Δ
Full model	**45.22**	**45.35**	37.01	–
w/o pretraining	35.76	33.27	23.13	−9.46
w/o the P2 level	43.69	44.94	32.01	−1.53
Contextual P2 only (w/o morphology path at P2)	43.73	39.73	34.57	−1.49
w/o morphology path	43.40	43.57	37.19	−1.83
Raw-only morphology path (w/o derivative branch)	43.12	42.01	36.92	−2.10
w/o homologous-pair interaction	41.11	39.71	32.23	−4.11
w/o channel self-attention	35.98	32.16	25.54	−9.24
w/o any cross-channel interaction	36.12	35.27	27.51	−9.10
Homologous-pair module randomly initialized at fine-tuning	42.00	42.44	35.20	−3.22
Homologous-pair module randomly initialized, matched continuation budget	43.36	44.60	37.46	−1.86
Masked-waveform pretraining only	35.41	32.05	22.21	−9.81
Spectral-token pretraining only	42.13	40.81	36.74	−3.09
Random rather than transient-biased temporal masks	42.43	42.55	36.40	−2.79
w/o channel masking in pretraining	44.32	44.91	**37.60**	−0.90
Channel-mask probability 0.2	42.73	40.60	34.66	−2.49

Table 16: Pyramid, head, and training ablations (%). Epi.: mean AP of sharp-wave and spike-and-slow-wave labels.

Variant	mAP@0.5	Epi. mAP	Abn. mAP	Δ
Full model	**45.22**	**45.35**	37.01	–
w/o sleep-stage auxiliary loss	44.32	44.49	35.16	−0.90
Allow all classes on P2	43.22	41.88	35.83	−2.00
w/o local temporal attention	41.88	38.73	35.31	−3.34
w/o top-down pyramid fusion	41.41	35.31	29.29	−3.81
Enable the additional P1 level	41.86	41.38	**38.00**	−3.36
w/o class-balanced sampling	44.08	43.48	34.25	−1.14
w/o temporal-IoU loss	41.41	39.48	32.32	−3.81
w/o center sampling	39.17	40.74	33.74	−6.05
w/o training augmentation	41.51	40.71	33.47	−3.71
w/o channel dropout	42.27	40.41	33.86	−2.95
w/o channel permutation	42.30	40.64	33.33	−2.92
w/o temporal shift	42.64	42.09	34.52	−2.59
w/o amplitude scaling	43.97	41.29	35.84	−1.25
w/o additive noise	43.79	41.52	34.77	−1.43
Ignore truncated training events	41.97	40.33	35.10	−3.25
w/o additional TU training data	34.59	35.44	34.77	−10.64

with abnormal slow wave and abnormal fast activity. These are label-based groups, with the sharp-wave class also including vertex transients. Δ is measured in percentage points relative to the full WA-Detector's 45.22% mAP.

E.2 Sensitivity to the Temporal Matching Threshold

Figure 8 sweeps the temporal IoU threshold used to match predicted and annotated events, holding the checkpoint, decoding, and non-maximum suppression settings fixed. Macro mAP decreases from 49.16% at tIoU 0.3 to 45.22% at tIoU 0.5, then falls more steeply to 28.85% at tIoU 0.7 and 23.23% at tIoU 0.75.

The per-class curves suggest different error patterns. Eye-blink and lateral eye-movement artifacts are among the strongest classes at tIoU 0.5 (52.11% and 61.16%) but lose most of their AP under stricter matching (8.66% and 13.79% at tIoU 0.75). This decrease suggests sensitivity to temporal boundary alignment. Abnormal slow wave and abnormal fast activity behave differently: their AP changes only slightly between tIoU 0.3 and 0.5 (28.18% to 27.79% and 32.56% to 29.56%, respectively). Their low AP at relaxed matching thresholds suggests that detection errors remain a major limitation. High-frequency noise and spike-and-slow-wave complexes retain AP values of 54.61% and 38.79%, respectively, at tIoU 0.75.

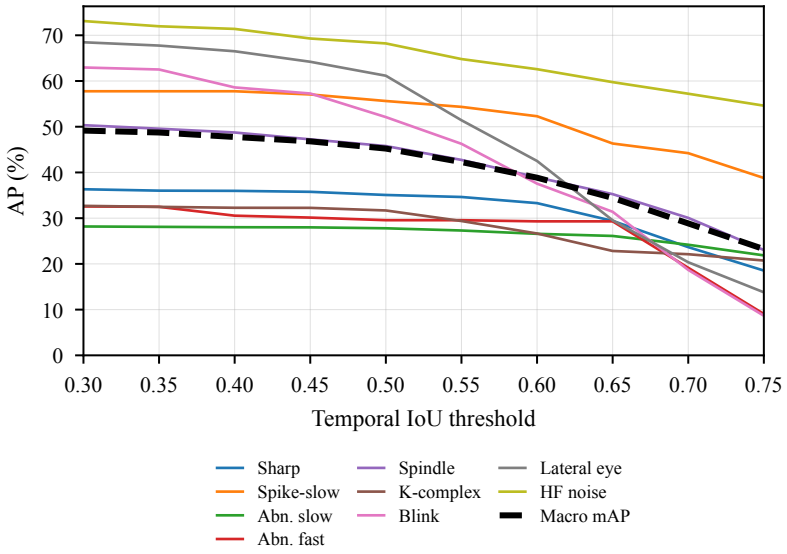


Figure 8: WA-Detector test AP versus tIoU on WA-Bench. Solid lines: per-class AP; dashed: macro mAP.

E.3 Event-Level Confusion Analysis

Figure 9 summarizes event-level confusion at the reporting threshold. "Background" denotes unmatched events rather than a separate EEG waveform class. The Background column records annotated events without a matched prediction (misses), whereas the Background row records predictions without a matched annotation (false positives); true-negative background intervals are not counted. High-frequency noise has the highest fraction of correctly localized and classified annotations (80.3%). Abnormal slow and fast activity often lack a matched prediction (60.1% and 60.3%, respectively), while 29.9% of K-complexes are predicted as sleep spindles. Sharp waves and abnormal slow waves also account for the largest numbers of unmatched predictions.

E.4 Qualitative Detection Examples

WA-Detector performs channel-level recognition of nine waveform classes: sharp wave, spike-and-slow-wave complex, abnormal fast activity, abnormal slow wave, high-frequency noise, eye-blink artifact, lateral eye-movement artifact, K-complex, and sleep spindle. Figure 10 presents representative detection examples by comparing WA-Detector's predicted intervals with expert annotations. All examples shown are drawn from WA-Bench.

F Additional WA-Agent Results and Evaluation Details

F.1 Evidence Configurations and Scoring

Structured evidence and textual summaries. The structured-evidence WA-Agent baseline uses WA-Detector events, background measurements, and sharp-transient measurements returned by the planner-selected tools. A supplementary-summary condition adds a deterministic textual summary while retaining the structured finding table as the primary evidence source. Summary-priority conditions instead instruct the responder to prioritize the summary's explicit class, channel, and interval relationships and use structured events to refine them. We evaluate this instruction both as an addition to the full responder prompt and as a concise replacement prompt.

The summary combines the same detector's events with a separate deterministic artifact analysis and one-second FFT background analysis. Its frequency and amplitude definitions need not match

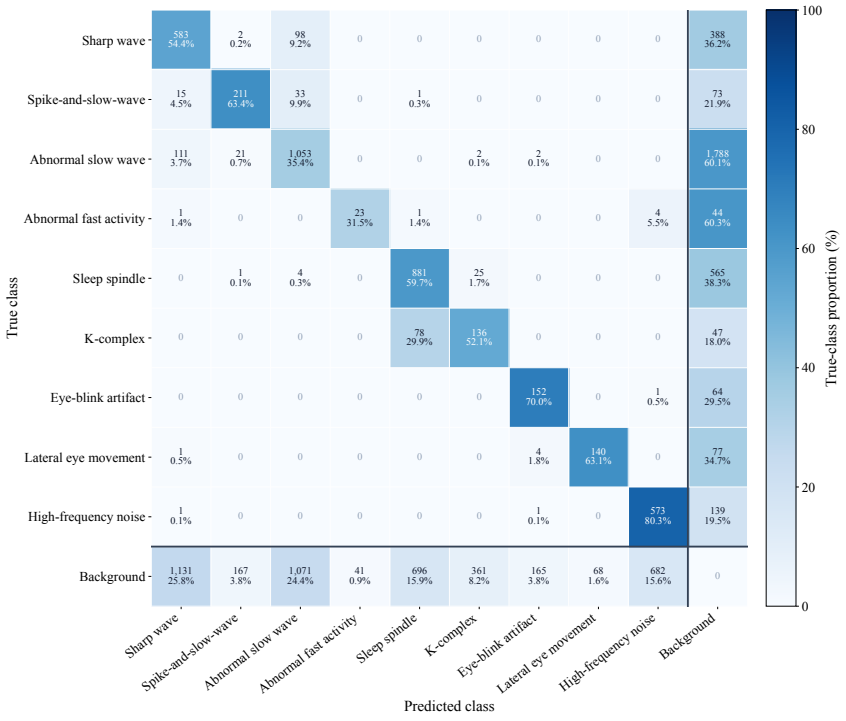


Figure 9: WA-Detector event-level confusion on WA-Bench's clinical test split at confidence threshold 0.202 and channel-matched tIoU 0.5. Matching is one-to-one, class-agnostic, and greedy in descending tIoU. Cells show counts and row-normalized percentages. The Background column contains unmatched annotations, and the Background row contains unmatched predictions; true-negative background intervals are not counted.

the interactive background tool's Welch/RMS measurements. Oracle event evidence substitutes annotated waveform boxes for detections. Oracle summaries also use annotated boxes; the condition combining detected events with an Oracle summary changes only the summary source relative to the corresponding summary-priority configuration. These annotation-assisted conditions have privileged test information and are diagnostic controls. Response generation under the summary and Oracle conditions reuses recorded planner routes, with each conversation continuing from that condition's own previous answers.

The planner and responder use DeepSeek-V4-Flash, with a maximum of 4,096 generated tokens per call and no LoRA adaptation. The default is three training-set demonstrations for each role. The detector uses decomposed eventness and a uniform reporting threshold of 0.202. The primary judge is DeepSeek-V4-Flash; the sensitivity analysis uses DeepSeek-V4-Pro. Both judges use the same three-level rubric prompt, a maximum of 768 output tokens, and per-turn conversation judging. In the demonstration-count experiments, we vary the number of examples given to the responder and reuse previously recorded planner calls. The supplementary-summary setting keeps the original responder prompt and provides the summary as additional evidence. The full-prompt priority setting adds instructions to prioritize the summary to the original prompt. The concise-priority setting instead uses a shorter replacement prompt.

F.2 Complete Configuration Results

Table 17 reports all 21 evaluated configurations. The main text emphasizes evidence presentation, tool availability, and routing; the remaining rows assess privileged summaries, demonstration count, and Oracle counterparts. D and C denote description and conversation. The concise summary-priority condition has an equal-task mean partial-credit score of 62.13%, compared with 56.28% for structured evidence alone.

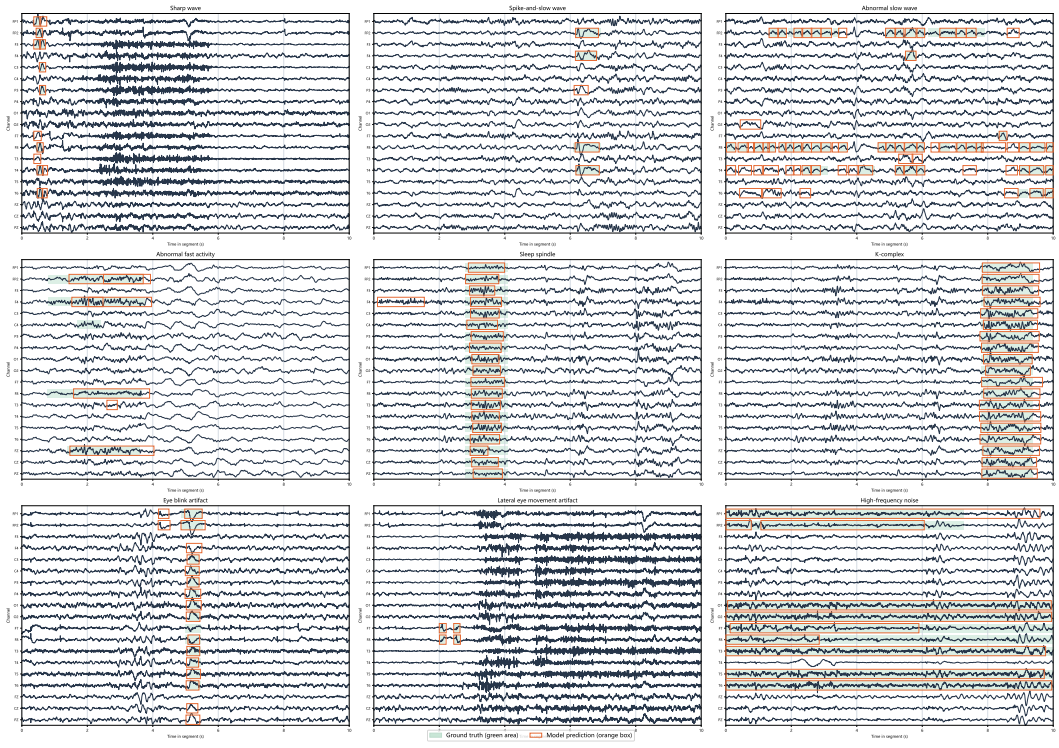


Figure 10: Qualitative WA-Detector examples on WA-Bench. Each panel highlights one waveform class in a 10-s EEG segment. Green shading marks reference annotations, and orange boxes mark predicted channel-level intervals.

Table 17: Complete WA-Agent results on WA-Bench (%). Det.: detected events; D/C: description/-conversation. Priority is concise unless marked full.

Configuration	Selection	D Strict	D Partial	C Strict	C Partial
Evidence and summary configuration					
Det., structured only	93.71	48.74	60.14	43.12	52.42
Det. + supplementary summary	94.29	48.43	60.05	45.33	54.84
Det. + priority summary, full prompt	93.14	49.28	61.48	47.84	56.60
Det. + priority summary	93.71	59.00	68.73	48.10	55.52
Oracle, structured only	97.71	71.63	79.75	77.58	83.10
Oracle + priority summary	98.29	91.88	94.84	93.66	94.86
Det. + Oracle priority summary[†]	97.71	88.27	92.16	82.43	86.07
Demonstration and routing controls					
Det. + priority, zero demonstrations	90.86	60.38	69.56	51.34	58.27
Det. + priority, one demonstration	90.86	59.81	69.45	51.03	58.29
Oracle + priority, one demonstration	94.86	92.71	95.55	93.27	94.54
Det., fixed three-tool routing	93.14	46.94	59.38	45.42	54.48
Oracle, fixed three-tool routing	98.29	71.08	79.08	78.61	83.89
Tool/evidence removal; structured only unless marked priority					
Det., w/o detection interface	86.29	29.18	34.65	14.53	17.05
Oracle, w/o detection interface	85.14	33.36	37.98	19.99	21.89
Det., w/o background tool	93.71	36.65	49.80	42.54	52.14
Oracle, w/o background tool	97.71	60.36	69.11	74.77	81.26
Det., w/o measurement tool	94.86	48.27	60.02	43.84	53.58
Oracle, w/o measurement tool	98.29	69.95	78.58	76.20	81.92
Det., w/o signal-quality evidence	93.71	46.21	59.50	41.99	51.52
Oracle, w/o signal-quality evidence	97.71	69.45	78.94	76.06	82.04
Det. + priority, w/o signal quality	94.29	56.51	67.88	49.57	57.17

[†]One failed description receives zero credit, and its corresponding selection response is counted as incorrect. All cases remain in the evaluation denominator.

Adding a supplementary summary gives a description partial-credit score of 60.05%, compared with 60.14% for structured evidence alone; how the responder is instructed to use the evidence matters alongside the evidence itself. With detected evidence, removing the measurement tool or signal-quality evidence produces comparatively small, task-dependent changes in the rubric scores (Table 17). Removing the detection interface also removes its signal-quality evidence, but detector-based sharp-transient measurement remains available. Some controls change multiple evidence sources and therefore do not isolate individual component contributions.

F.3 Judge Sensitivity

DeepSeek-V4-Pro re-scores the same outputs for the four principal detector/Oracle and structured/priority conditions using the same rubric prompt and per-turn context. It assigns higher absolute scores, but preserves their ordering for both description and conversation (Table 18).

Table 18: Judge sensitivity: description (D) and conversation (C) partial-credit scores (%).

Configuration	D Flash	D Pro	C Flash	C Pro
Det., structured	60.14	70.74	52.42	63.25
Det. + priority	68.73	79.92	55.52	66.30
Oracle, structured	79.75	86.06	83.10	88.65
Oracle + priority	94.84	98.70	94.86	96.03

F.4 Routing and Rendering Diagnostics

We check whether required tools are called with valid arguments and execute successfully, counting extra calls separately. The structured-evidence baseline passes all three checks on 860 requests (Table 19). It makes 1,408 tool calls (1.637 per request), with extra calls on 149 requests.

Routing reference. Required tools are defined from the question stem and planner instructions, not from multiple-choice distractors. Selection requires background analysis for the posterior-rhythm item and detection otherwise. Description requires detection, background analysis, and measurement. Background-only conversation questions require background analysis; other questions require detection. Argument checks require the correct study, schema-valid arguments, and the full case interval or an interval explicitly requested in the input. Arbitrary contained subintervals are not accepted; rendering also checks overlay mode. Extra calls are the multiset difference between called and required tool names, including repeated calls.

Call composition. For structured detector evidence, selection makes 365 calls, description 525, and conversation 518. Across these requests, detection is invoked 846 times, background analysis 293 times, and measurement 269 times. Summary/Oracle replays reuse this call pattern except where a generation fails and do not constitute independent planner decisions. Fixed routing invokes each of the three analysis tools 860 times.

Plot delivery and exact-call matching. The WA-Bench visualization task covers raw-to-overlay, raw-to-discussion, discussion-to-overlay, and overlay-to-discussion transitions. Each second turn receives the actual first-turn answer. Both additional background calls occur on discussion turns; all required detection and rendering calls execute successfully. The exact-call metric fails those two turns, while required-tool coverage passes them (Table 20). All image turns validate the requested study, interval, and overlay mode, expose the URL in the tool result and response, and retrieve a valid PNG. Byte equality with reference raw plots passes on 15/15 raw-image turns as a diagnostic. Annotated reference plots use privileged boxes and are not pixel-matched to the detector overlays. The evaluation does not establish the clinical validity of the overlays.

F.5 Supplementary NLP Metrics

Protocol. We report ROUGE-1 and ROUGE-L F1 (Lin, 2004) for every configuration in Table 21. All configurations share one manually reviewed synthetic silver reference per response; these references are separate from the evidence summaries and rubric criteria. Before scoring, fixed safety

Table 19: Tool-use reliability for the structured baseline and a separate plotting evaluation.

Task	Requests	Coverage	Arguments	Execution	Extra-call requests
Selection	175	175/175	175/175	175/175	141/175
Description	175	175/175	175/175	175/175	0/175
Conversation	510	510/510	510/510	510/510	8/510
Plotting/discussion	60	60/60	60/60	60/60	2/60

Table 20: Plotting reliability: coverage permits extra calls; exact-call matching does not.

Criterion	Passed / total	Rate (%)
Successful API response	60/60	100.00
Required-tool coverage	60/60	100.00
Valid arguments and successful required calls	60/60	100.00
Exact-call pass	58/60	96.67
Valid PNG and artifact delivery	37/37	100.00
Second-turn exact-call pass	30/30	100.00
Complete two-turn exact-call pass	28/30	93.33
Raw-image byte match (diagnostic)	15/15	100.00

notices and appended tool-background measurement blocks are removed from both prediction and reference, and whitespace is collapsed. The EEG-aware tokenizer lowercases text, retains alphanumeric electrode labels and numeric tokens, normalizes the two micro-symbol forms, and does not apply stemming. ROUGE-1 uses clipped unigram overlap; ROUGE-L uses the longest common subsequence within each response, not sentence-union ROUGE-Lsum.

Description scores average over 175 cases. Conversation scores first average response scores within each case and then average equally over the 175 cases, covering all 510 reference turns, including the turn without a source rubric. Failed or missing responses receive zero ROUGE scores and remain in the denominator. These conversation scores must not be mixed with response-pooled aggregates, which give more weight to longer conversations. Under this tokenizer, ROUGE-1 F1 is equivalent to token-overlap F1 and is reported only once.

Table 22 additionally reports BLEU-1/2/4 (Papineni et al., 2002) with add-one smoothing of the clipped n-gram precision counts and the usual brevity penalty. BLEU pools counts over the corpus separately for description and conversation; it is not a case-macro average. All displayed lexical scores are multiplied by 100.

Limited alignment with the task. Lexical overlap can correlate poorly with human judgments of generated text (Liu et al., 2023). These lexical metrics are poorly aligned with the primary objective of evidence-grounded EEG interpretation. They measure similarity to a particular reference wording, rather than whether an answer correctly associates a waveform with its electrode, time interval, laterality, or measured amplitude. A wrong channel, interval, or negation can leave most words unchanged, while a correct paraphrase, a different ordering of findings, or additional supported detail can reduce overlap. A single synthetic reference also need not exhaust all valid descriptions. Removing boilerplate reduces one source of overlap but does not make the remaining score a measure of clinical correctness.

The observed results illustrate this mismatch. Removing the detection interface raises description ROUGE-1 from 16.50 to 18.51 and ROUGE-L from 9.66 to 10.73, while the partial-credit rubric falls from 60.14% to 34.65%. For Oracle evidence, prioritizing the summary raises the conversation rubric from 83.10% to 94.86%, yet ROUGE-1 falls from 20.84 to 18.70 and ROUGE-L from 14.07 to 12.64. Thus higher lexical overlap does not consistently track better rubric performance in this evaluation. We retain these metrics as supplementary measures of text similarity and use the task-specific rubric as the primary measure of answer quality.

Table 21: Case-macro ROUGE F1 (0–100): R-1 = ROUGE-1; R-L = ROUGE-L.

	Description		Conversation	
Configuration	R-1	R-L	R-1	R-L
Evidence and summary configuration				
Det., structured only	16.50	9.66	16.65	11.32
Det. + supplementary summary	16.83	11.07	16.82	11.51
Det. + priority summary, full prompt	16.09	10.52	16.90	11.56
Det. + priority summary	18.42	11.25	16.62	11.25
Oracle, structured only	19.48	11.45	20.84	14.07
Oracle + priority summary	18.77	11.54	18.70	12.64
Det. + Oracle priority summary	16.38	10.14	15.84	10.75
Demonstration and routing controls				
Det. + priority, zero demonstrations	15.52	9.42	13.86	9.41
Det. + priority, one demonstration	17.29	10.16	16.41	10.93
Oracle + priority, one demonstration	18.60	11.67	18.17	12.20
Det., fixed three-tool routing	16.31	9.61	14.85	10.07
Oracle, fixed three-tool routing	19.37	11.38	17.91	12.08
Tool/evidence removal; structured only unless marked priority				
Det., w/o detection interface	18.51	10.73	14.20	10.07
Oracle, w/o detection interface	22.25	12.70	15.24	10.77
Det., w/o background tool	18.66	11.28	16.64	11.33
Oracle, w/o background tool	21.78	13.18	22.14	14.95
Det., w/o measurement tool	18.17	10.64	16.94	11.47
Oracle, w/o measurement tool	22.56	13.49	22.23	14.93
Det., w/o signal-quality evidence	16.79	9.96	17.42	11.66
Oracle, w/o signal-quality evidence	19.83	11.76	22.05	14.81
Det. + priority, w/o signal quality	18.93	11.48	17.48	11.82

Table 22: Add-one-smoothed corpus BLEU (0–100). B-1/2/4 denote BLEU-1/2/4.

	Description			Conversation		
Configuration	B-1	B-2	B-4	B-1	B-2	B-4
Evidence and summary configuration						
Det., structured only	9.63	5.07	1.69	10.36	5.26	1.73
Det. + supplementary summary	9.96	5.57	2.17	10.24	5.42	1.87
Det. + priority summary, full prompt	9.44	5.25	2.05	10.09	5.29	1.84
Det. + priority summary	10.59	6.10	2.38	9.48	5.09	1.81
Oracle, structured only	11.85	6.66	2.40	13.09	7.40	2.77
Oracle + priority summary	10.37	6.25	2.54	10.46	6.28	2.59
Det. + Oracle priority summary	9.17	5.36	1.99	8.78	5.14	2.09
Demonstration and routing controls						
Det. + priority, zero demonstrations	8.48	4.95	1.90	7.44	4.00	1.44
Det. + priority, one demonstration	9.72	5.49	2.01	9.38	4.97	1.76
Oracle + priority, one demonstration	10.14	6.20	2.56	10.26	6.16	2.52
Det., fixed three-tool routing	9.55	5.03	1.67	8.57	4.32	1.37
Oracle, fixed three-tool routing	11.77	6.65	2.40	10.51	5.78	2.09
Tool/evidence removal; structured only unless marked priority						
Det., w/o detection interface	11.28	5.37	1.89	10.02	4.71	1.62
Oracle, w/o detection interface	14.74	7.68	3.00	11.65	6.02	2.16
Det., w/o background tool	11.19	5.74	2.02	10.37	5.26	1.73
Oracle, w/o background tool	13.75	7.54	2.82	14.29	8.08	3.09
Det., w/o measurement tool	10.98	5.86	1.97	10.57	5.33	1.75
Oracle, w/o measurement tool	14.21	8.19	3.19	14.47	8.13	3.02
Det., w/o signal-quality evidence	9.89	5.16	1.72	11.00	5.53	1.81
Oracle, w/o signal-quality evidence	12.08	6.84	2.62	14.23	7.99	3.01
Det. + priority, w/o signal quality	10.86	6.09	2.28	9.99	5.38	1.93

G Qualitative Results

G.1 WA-Agent

We show three WA-Bench cases. WA-Agent responses are excerpts, with [. . .] marking omissions. EEG overlays show reference annotations as green shading and detector predictions as red boxes. Table 23 summarizes the outcomes.

Table 23: Recorded rubric outcomes for the selected cases. Entries count pass/partial/fail judgments for complete responses; conversation counts pool the criteria from all scored turns.

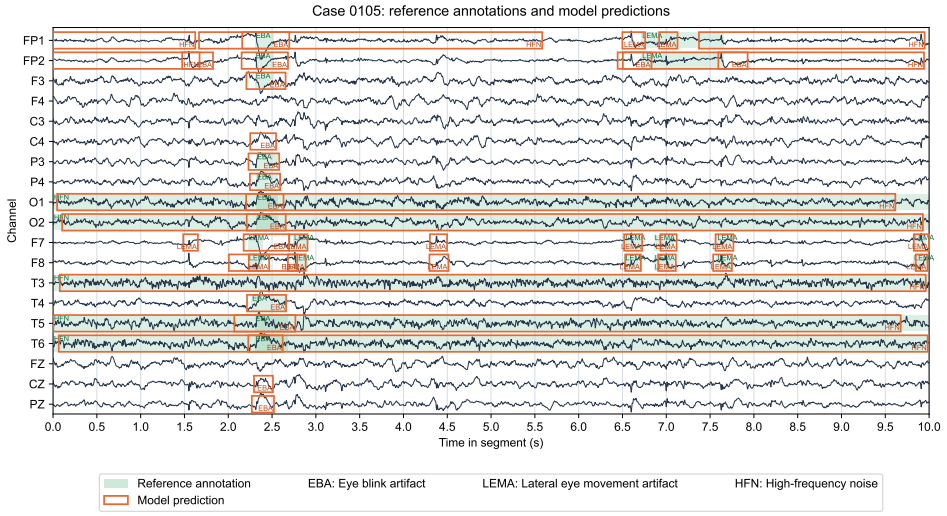
Case	Description	Conversation	Main observation
0105	5/0/0	2/0/1	Artifact recognition, with one conversation distribution criterion missed.
0084	4/0/0	6/0/0	Main waveform classes and their right-sided emphasis are captured.
0160	1/2/3	0/2/4	Spurious left-sided sharp-wave predictions distort the reported distribution.

Implications. These cases make the detector-to-answer relationship inspectable, showing both successful waveform interpretation and the propagation of spurious detections into generated answers. Grounding provides a traceable basis for an answer, while its correctness still depends on the evidence supplied and how the responder uses it.

Positive example: artifact recognition. Case 0105 identifies an artifact-dominated segment without treating the transients as definite epileptiform discharges (Table 24). All description criteria pass and the selection answer is correct. The first conversation turn fails its high-frequency-noise distribution criterion, while the subsequent eye-movement and gaze-artifact criteria pass.

Table 24: Case 0105: description, selection, and conversation.

Source Test case 0105, recording 39, segment 12 (0–10 s).



Description

User	Please provide a formal clinical description of this 10-second EEG segment.
WA-Agent	This 10-second segment uses the standard 10-20 montage with an average reference. [. . .] High-frequency noise: O1 0.0468–9.6147 s; T6 0.065–9.9835 s; T3 0.0771–9.9618 s; plus additional frontal/polar involvement (FP1, FP2, F7) reported across the segment. [. . .] Synthesis: this is a fair-quality, artifact-rich segment with prominent high-frequency noise and eye-related artifacts, no detected epileptiform discharges, and no reported sleep graphoelements. [. . .]
Rubric outcomes	**Q1: Pass.** Recording uses the standard 10-20 system with an average reference. **Q2: Pass.** High-frequency noise over bilateral temporal and occipital regions throughout, with additional frontal noise at FP2 and F7. **Q3: Pass.** Eyeblink artifact over the frontopolar region. **Q4: Pass.** Eye-movement artifact involving the temporal and posterior regions. **Q5: Pass.** Lateral eye-movement artifact over the temporal regions.

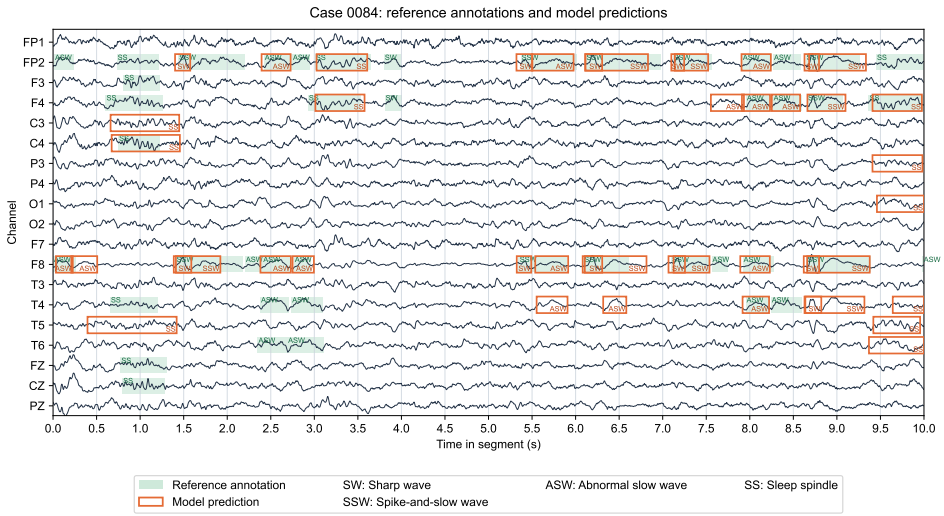
Table 24 (continued)

Selection	
User	What is the most appropriate electrographic interpretation of this segment? Options: A) Generalized epileptiform activity with bifrontal emphasis B) No definite epileptiform discharges; frequent eye movement and muscle artifacts C) Left temporal lobe epileptiform discharges D) Diffuse background slowing consistent with encephalopathy Select the best option and briefly explain the evidence.
WA-Agent	**Best option: B) No definite epileptiform discharges; frequent eye movement and muscle artifacts.** [. . .]
Answer Key	B) No definite epileptiform discharges; frequent eye movement and muscle artifacts
Selection result	Correct (B).
Conversation	
User (turn 1)	What is the distribution of the main waveform activity in this segment?
WA-Agent	The main activity in this 10-second segment is artifactual rather than physiological. The tool returned 44 events, all classified as artifact (0 waveform events, 0 epileptiform candidates), with the following distribution: **High-frequency noise (10 events)** - O1: 0.0468–9.6147 s - T6: 0.065–9.9835 s - T3: 0.0771–9.9618 s - FP1: 3 events; FP2: 2 events; O2: 1 event; T5: 1 event (per-channel counts; individual intervals not all listed in the returned subset) [. . .]
Rubric outcomes	**Q1: Fail.** The description identifies high-frequency noise as the main activity over bilateral temporal and occipital regions throughout, with additional noise over right frontopolar and left inferior frontal regions.
User (turn 2)	Where is the eye-movement artifact distributed?
WA-Agent	[. . .] Eye blink artifact is maximal at the frontopolar leads (FP2 with 4 events, FP1 with 1) and spreads broadly across frontal, temporal, occipital, and parietal channels; its overall time range is 1.4737–7.9285 s. [. . .]
Rubric outcomes	**Q1: Pass.** The description identifies eye-movement artifact around 2.3–2.6 s involving the frontopolar region and bilateral temporal, parietal, and occipital regions.
User (turn 3)	What about the lateral eye-movement/gaze artifact?
WA-Agent	[. . .] The class is confined to F7 (7 events), F8 (7 events), and FP1 (2 events) — i.e., frontotemporal, with the two frontal-temporal leads carrying the bulk of it. [. . .]
Rubric outcomes	**Q1: Pass.** The description identifies intermittent lateral eye-movement artifact involving bilateral inferior frontal/temporal regions, with additional frontopolar involvement later in the segment.

Positive example: mixed-waveform interpretation. Case `0084` captures the main waveform classes and their right-sided emphasis (Table 25). The selection answer is correct, and all description and conversation criteria pass.

Table 25: Case 0084: description, selection, and conversation.

Source	Test case 0084, recording 37, segment 44 (0–10 s).



Description

User	Please provide a formal clinical description of this 10-second EEG segment.
WA-Agent	[. . .] The record shows frequent sharp transients and spike-and-slow-wave complexes with a right frontotemporal emphasis (F8, FP2, T4, F4), together with abnormal slow-wave activity in the same regions, alongside sleep spindles indicating a sleep or drowsy state. [. . .]
Rubric outcomes	**Q1: Pass.** The description identifies sleep spindles over bilateral frontal and central regions and the right temporal region, occurring intermittently. **Q2: Pass.** The description identifies spike-and-slow-wave complexes over the right frontopolar, right frontal, and right anterior temporal regions, occurring intermittently. **Q3: Pass.** The description identifies slow waves over the right frontopolar, right frontal, and right temporal regions, occurring intermittently. **Q4: Pass.** The description identifies sharp waves over the right frontopolar and right frontal regions.

Selection

User	What is the most likely localization of the epileptiform discharges seen in this segment? Options: A) Generalized with no consistent field B) Bilateral posterior quadrants C) Left centro-parietal region D) Right frontal/anterior-temporal region (Fp2, F4, F8) Select the best option and briefly explain the evidence.
WA-Agent	**Answer: D) Right frontal/anterior-temporal region (Fp2, F4, F8).** [. . .]
Answer Key	D) Right frontal/anterior-temporal region (Fp2, F4, F8)
Selection result	Correct (D).

Conversation

User (turn 1)	What are the main waveforms in this 10-second segment and how are they distributed?
WA-Agent	[. . .] The distribution is notable for a right-sided emphasis: sharp waves, spike-and-slow-wave complexes, and abnormal slow waves cluster over F8, FP2, F4, and T4, with the most prominent spike-and-slow-wave activity at F8 (8.659–9.3771 s) and T4 (8.6379–9.3162 s). Sleep spindles are more widely distributed across frontal, central, parietal, occipital, and temporal channels, appearing both early (around 0.4–1.4 s) and late (around 9.4–10.0 s). The background is low-amplitude and mixed in frequency without a clear dominant posterior rhythm. [. . .]

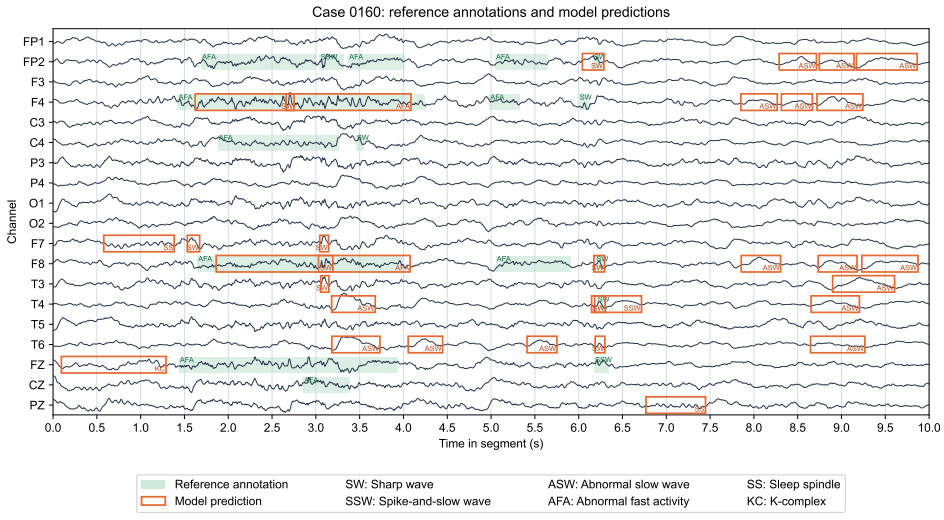
Table 25 (continued)

Rubric outcomes	**Q1: Pass.** The description identifies sleep spindles distributed over the frontopolar, bilateral frontal, central, and right temporal regions. **Q2: Pass.** The description identifies spike-and-slow-wave complexes over the right frontopolar and right frontal regions. **Q3: Pass.** The description identifies slow-wave activity over the right frontopolar, right frontal, and right temporal regions. **Q4: Pass.** The description identifies sharps over the right frontopolar and right frontal regions.
User (turn 2)	What is the distribution of the spike-and-slow-wave complexes?
WA-Agent	[. . .] In total, 9 spike-and-slow-wave events were detected across 4 channels (F8: 4, FP2: 3, F4: 1, T4: 1), spanning 1.4107–9.3771 s. The distribution is right-sided and predominantly frontotemporal, with the greatest concentration over F8 and FP2. [. . .]
Rubric outcomes	**Q1: Pass.** The description identifies spike-and-slow-wave complexes distributed over the right frontopolar and right frontal regions.
User (turn 3)	What about the slow-wave activity?
WA-Agent	[. . .] In total, 15 abnormal slow-wave events were detected across 4 channels (F8: 6, FP2: 3, F4: 3, T4: 3), spanning 0.0331–8.5779 s. The distribution is right-sided and predominantly frontal/frontotemporal, with the greatest number over F8. [. . .]
Rubric outcomes	**Q1: Pass.** The description identifies slow-wave activity distributed over the right frontopolar, right frontal, and right temporal regions.

Negative example: sharp-wave misidentification. In case `0160`, the detector adds sharp-wave predictions at left-sided F7 and T3 and misses the annotated C4 event. The answer consequently describes a bilateral distribution, whereas the reference sharp-wave labels are right-sided. The sharp-wave criteria fail in both description and the focused conversation turn (Table 26).

Table 26: Case `0160`: sharp-wave criteria in description and conversation.

Source	Test case `0160`, NREM III recording, segment 1 (0–10 s).



Detector evidence	Sharp-wave predictions include F7 at 1.5325–1.6732 s and 3.0448–3.1452 s, and T3 at 3.0549–3.1490 s, without corresponding reference sharp-wave labels. No sharp wave is predicted at C4, where a reference event occurs at 3.4590–3.5420 s.
Description	
User	Please provide a formal clinical description of this 10-second EEG segment.
WA-Agent	[. . .] **Sharp transients (9 events).** - F7: 1.5325–1.6732 s; 3.0448–3.1452 s. [. . .] - T3: 3.0549–3.149 s. [. . .]

Table 26 (continued)

Rubric outcomes	**Q4: Fail.** The description identifies sharp waves over the right frontal, temporal, and central regions, including right frontotemporal sharp waves.
Conversation	
User (turn 2)	Where are the sharp waves distributed?
WA-Agent	[. . .] Because sharp waves appear on both left and right sides, the distribution is bilateral, but the two sides are not equally represented at the same times. [. . .]
Rubric outcomes	**Q1: Fail.** Identifies sharp waves over the right frontal and frontotemporal regions, and a separate sharp wave at C4.

H WA-Agent Planner and Responder Model Comparisons

H.1 Configurations and Scoring

The matrix evaluates nine planner/responder pairs on WA-Bench using DeepSeek-V4-Flash, DeepSeek-V4-Pro, Qwen3-8B, and Qwen3-32B (Yang et al., 2025). Flash and Pro abbreviate the DeepSeek models in the tables. Each pair is evaluated with detected or Oracle evidence and with structured-only or concise summary-priority responses, giving 36 task configurations. Each configuration contains 175 selection cases, 175 descriptions, and 175 conversations. Description and conversation use the same 801 and 749 rubric items as the main experiment, respectively.

The shared WA-Detector checkpoint has a reporting threshold of 0.202. Planner and responder prompting uses three training-set demonstrations, without responder LoRA adaptation, and the generation limit is 4,096 tokens per call. Qwen models are served locally, while the DeepSeek models use the API. The judge is DeepSeek-V4-Flash with the same three-level prompt, per-turn conversation context, and a 768-token output limit. Table 27 shows the generation settings that differ across model configurations.

Table 27: Generation settings that differ across model configurations.

Condition	Setting
Structured Flash/Qwen3-8B and Flash/Pro	Completed outputs are retained across resumed runs; prompt-file versions differ between the initial and resumed runs.
Qwen summary-priority recovery	Thinking disabled; temperature 0.7; top-p 0.8; top-k 20; minimum-p 0; output limit 4,096 tokens. Applies to both evidence conditions with a Qwen3-32B responder and either Qwen planner, and to detected evidence with Qwen3-32B/Qwen3-8B.
Qwen3-32B responder history	Up to 12 messages and 16,000 total characters; per-message limits of 16,000 assistant and 6,000 user characters.

H.2 Complete Task Results

Table 28 reports every model/evidence/response configuration. All tested pairs have higher description and conversation partial-credit scores in the summary-priority conditions with detected evidence, although selection accuracy does not improve uniformly. The Oracle conditions retain differences between responder backends even when the waveform evidence is supplied by annotations.

H.3 Tool Use and Visualization

Table 29 compares tool-use reliability and call counts across planner–responder pairs. Required-tool coverage ranges from 98.95% to 100% across all pairs. The Qwen-only pairs use fewer calls per request and have fewer requests with extra calls than Flash/Flash, although their required-tool coverage is slightly lower.

The separate visualization evaluation uses 30 free-running two-turn cases for every model pair (Table 30). Every configuration produces successful API responses for all 60 turns and delivers valid

Table 28: Complete model matrix (%). Rubrics are strict/partial; Flash/Pro denote DeepSeek-V4-Flash/Pro.

Planner	Responder	Selection	Description	Conversation
Detected events, structured only				
Flash	Flash	93.71	48.74 / 60.14	43.12 / 52.42
Qwen3-8B	Flash	92.57	47.77 / 60.04	41.74 / 51.52
Flash	Qwen3-8B	82.29	31.12 / 46.22	31.54 / 46.02
Pro	Flash	93.14	47.40 / 59.38	41.77 / 51.83
Flash	Pro	94.29	43.58 / 57.86	41.01 / 52.39
Pro	Pro	95.43	45.00 / 58.07	37.75 / 49.74
Qwen3-32B	Qwen3-32B	88.57	34.83 / 49.10	32.99 / 46.34
Qwen3-8B	Qwen3-32B	90.86	35.89 / 49.66	33.46 / 46.52
Qwen3-32B	Qwen3-8B	82.86	30.85 / 45.88	31.49 / 46.68
Detected events, summary priority				
Flash	Flash	93.71	59.00 / 68.73	48.10 / 55.52
Qwen3-8B	Flash	93.14	57.17 / 67.73	50.59 / 58.01
Flash	Qwen3-8B	78.29	45.33 / 59.65	42.34 / 54.58
Pro	Flash	93.14	57.37 / 67.91	50.70 / 58.46
Flash	Pro	93.71	55.52 / 65.85	49.24 / 57.75
Pro	Pro	93.14	55.38 / 65.92	50.00 / 58.40
Qwen3-32B	Qwen3-32B	87.43	51.51 / 63.80	45.84 / 56.07
Qwen3-8B	Qwen3-32B	86.86	52.15 / 64.32	44.65 / 55.30
Qwen3-32B	Qwen3-8B	74.86	45.57 / 60.20	41.52 / 53.78
Oracle evidence, structured only				
Flash	Flash	97.71	71.63 / 79.75	77.58 / 83.10
Qwen3-8B	Flash	96.00	70.57 / 78.90	78.24 / 83.31
Flash	Qwen3-8B	89.71	48.84 / 63.08	55.73 / 69.07
Pro	Flash	98.29	71.03 / 79.48	76.89 / 83.08
Flash	Pro	97.71	66.70 / 76.82	70.55 / 79.64
Pro	Pro	97.71	68.30 / 77.55	72.81 / 80.57
Qwen3-32B	Qwen3-32B	92.57	53.73 / 67.03	60.11 / 71.68
Qwen3-8B	Qwen3-32B	93.71	54.65 / 67.90	62.60 / 73.57
Qwen3-32B	Qwen3-8B	85.71	44.74 / 58.90	57.52 / 70.96
Oracle evidence, summary priority				
Flash	Flash	98.29	91.88 / 94.84	93.66 / 94.86
Qwen3-8B	Flash	97.71	91.39 / 94.76	92.47 / 93.94
Flash	Qwen3-8B	78.29	66.46 / 79.44	77.43 / 84.39
Pro	Flash	97.14	91.80 / 95.11	94.84 / 95.77
Flash	Pro	98.86	88.23 / 92.98	89.37 / 92.10
Pro	Pro	98.29	90.91 / 94.52	89.15 / 92.04
Qwen3-32B	Qwen3-32B	89.14	79.40 / 86.40	83.08 / 87.22
Qwen3-8B	Qwen3-32B	92.57	81.16 / 87.47	84.66 / 89.14
Qwen3-32B	Qwen3-8B	81.71	68.65 / 80.27	77.52 / 84.00

Table 29: Live structured-evidence routing. Calls: calls/request; Extra req.: requests with extra calls.

Planner	Responder	Coverage	Calls	Extra req.	All calls OK
Flash	Flash	860/860	1.637	149	860/860
Qwen3-8B	Flash	857/860	1.535	97	859/860
Flash	Qwen3-8B	858/860	1.662	143	860/860
Pro	Flash	851/860	1.653	156	860/860
Flash	Pro	860/860	1.658	138	860/860
Pro	Pro	853/860	1.656	155	860/860
Qwen3-32B	Qwen3-32B	854/860	1.447	30	860/860
Qwen3-8B	Qwen3-32B	853/860	1.529	96	859/860
Qwen3-32B	Qwen3-8B	854/860	1.448	31	860/860

PNGs for all 37 rendering turns. Exact-call matching additionally requires the expected call, exact arguments, and successful execution. The Qwen3-32B/Qwen3-32B pair selects the expected tool on every turn but has six argument mismatches, illustrating the distinction between file delivery and exact instruction following.

Table 30: Visualization by model pair. Exact turns include argument and execution checks; latency is seconds/turn.

Planner	Responder	Exact turns	Full cases	PNGs	Mean latency
Flash	Flash	58/60	28/30	37/37	2.88
Qwen3-8B	Flash	57/60	27/30	37/37	3.67
Flash	Qwen3-8B	58/60	28/30	37/37	7.85
Pro	Flash	59/60	29/30	37/37	3.84
Flash	Pro	58/60	28/30	37/37	4.77
Pro	Pro	59/60	29/30	37/37	5.90
Qwen3-32B	Qwen3-32B	54/60	24/30	37/37	19.77
Qwen3-8B	Qwen3-32B	56/60	26/30	37/37	14.02
Qwen3-32B	Qwen3-8B	58/60	28/30	37/37	9.75

H.4 Recorded Latency

Table 31 gives the recorded mean request latency for every task configuration. Structured detector runs include live planning and tool execution; Oracle runs replay routes, and summary-priority runs reuse recorded evidence. The columns therefore have different execution scopes. Local Qwen serving and remote API execution also differ in deployment and concurrency, so these values describe the recorded runs rather than a normalized model-speed benchmark.

Table 31: Recorded mean request latency (s). Execution scopes differ between conditions.

		Structured only		Summary priority	
Planner	Responder	Detected	Oracle	Detected	Oracle
Flash	Flash	120.81	98.20	3.56	3.72
Qwen3-8B	Flash	151.80	101.26	3.59	3.80
Flash	Qwen3-8B	282.71	181.34	22.08	21.68
Pro	Flash	156.13	101.66	3.74	3.83
Flash	Pro	152.39	84.44	8.71	8.40
Pro	Pro	178.17	115.59	8.32	7.72
Qwen3-32B	Qwen3-32B	852.05	672.40	48.05	44.01
Qwen3-8B	Qwen3-32B	867.23	690.93	47.97	42.80
Qwen3-32B	Qwen3-8B	538.29	307.89	28.56	25.69

I CerebraGloss Comparisons

I.1 Models, Label Schemas, and Evaluation Settings

The detector comparison concerns CerebraGloss-YOLO, the waveform detector used in CerebraGloss's training-data pipeline, whereas the language comparison concerns the image-based CerebraGloss model (Gu et al., 2026). WA-Agent queries WA-Detector and signal-analysis tools at inference. The two comparisons therefore evaluate waveform localization and complete question-answering systems, respectively.

Dataset-specific detectors. On WA-Bench, both detectors use the waveform classes in Section 3, with spikes included in the sharp-wave class and separate abnormal-slow and abnormal-fast classes. The adapted CerebraGloss-YOLO uses the released model architecture trained for this label set.

On CerebraGloss-Bench, WA-Detector uses a separately trained checkpoint with the benchmark's native classes: sharp waves, spikes, spike-and-slow-wave complexes, sleep spindles, K-complexes,

blinks, positive and negative lateral eye movements, and high-frequency noise. Both the class vocabulary and the learned weights therefore differ from those of the model evaluated on our data. Evaluation matches predictions and annotations one-to-one within the same segment, class, and channel at temporal IoU 0.5. With a score threshold of 0.01, all-points interpolated AP yields 44.39% mAP; 11-point interpolation yields 43.86%. There are 1,793 annotated events and 6,872 predictions. The 40.95% CerebraGloss-YOLO baseline is quoted from the original paper's Table 2; its class-wise results below are quoted from Table 6. The comparisons retain these reported evaluation protocols.

WA-Bench language evaluation. CerebraGloss is evaluated on the same 175 WA-Bench source segments as WA-Agent. Description and conversation use exactly the same rubric items and the same three-level DeepSeek-V4-Flash judge prompt. Conversation is scored per turn, followed by case-level aggregation. These are comparisons of the evaluated system configurations rather than a controlled substitution of only the language model.

CerebraGloss-Bench language evaluation. On CerebraGloss-Bench, WA-Agent uses the benchmark-native WA-Detector with a reporting threshold of 0.01 and one external synthetic demonstration. All language results in Table 34 use the supplied benchmark outputs and a common DeepSeek-V4-Flash conversation judge. The judge compares answers with the benchmark reference and context in terms of helpfulness, relevance, accuracy, and detail, returning a final score out of 10. It uses temperature 0.2 and a 5,000-token output limit. These are re-scored results, distinct from the original paper's GPT-5-judged scores. Summary is evaluated with mean ROUGE-1 F1; the conversation score is not the case-specific EEG rubric used on our 175-case WA-Bench. Each task covers 90 cases.

I.2 Complete Detector Comparisons

Table 32: Per-class test AP@0.5 on WA-Bench (%).

Waveform class	CerebraGloss-YOLO	WA-Detector
Sharp wave	26.50	**35.07**
Spike-and-slow wave	37.71	**55.62**
Abnormal slow wave	**29.34**	27.79
Abnormal fast activity	0.10	**29.56**
Sleep spindle	38.12	**45.78**
K-complex	5.53	**31.68**
Eye-blink artifact	**56.31**	52.11
Lateral eye movement	50.01	**61.16**
High-frequency noise	67.43	**68.22**
Macro mAP	34.56	**45.22**

Table 33: CerebraGloss-Bench detection (%). Published AP values are rounded.

Waveform class	Published YOLO AP[†]	WA-Detector AP
Sharp wave	**63.00**	51.82
Spike	19.00	**21.62**
Spike-and-slow wave	**0.00**	**0.00**
Sleep spindle	**71.00**	67.90
K-complex	47.00	**66.92**
Eye-blink artifact	**76.00**	69.79
Lateral eye movement (+)	4.00	**21.65**
Lateral eye movement (-)	15.00	**29.89**
High-frequency noise	**75.00**	69.96
Macro mAP	40.95	**44.39**

[†]CerebraGloss-YOLO results are taken from Gu et al. (2026): per-class AP from Table 6 and macro mAP from Table 2.

The benchmark class-wise AP values published for CerebraGloss-YOLO are rounded to two decimal places on the 0–1 scale; their mean need not reproduce the separately reported 40.95% mAP exactly. Both detectors have zero AP for spike-and-slow-wave complexes on this benchmark, while their relative performance varies across the other classes.

I.3 Complete Language Results

Table 34: CerebraGloss-Bench language results. Selection and ROUGE-1 are percentages; conversation is scored out of 10.

System	Parsed	Selection: all / parsed	Summary R-1	Conversation
CerebraGloss‡	90/90	**80.00 / 80.00**	**44.19**	4.10
GPT-5‡	90/90	70.00 / 70.00	35.90	3.80
Gemini 2.5 Pro‡	90/90	64.44 / 64.44	37.02	3.42
LLaVA-Med‡	3/90	2.22 / 66.67	8.87	1.91
BioMedGPT‡	4/90	1.11 / 25.00	11.82	1.23
CerebraGloss w/o LLM‡	0/90	0.00 / –	9.02	2.41
WA-Agent, no summary	90/90	70.00 / 70.00	25.76	4.27
WA-Agent, compact summary	90/90	65.56 / 65.56	28.20	**4.51**
WA-Agent, expanded summary	90/90	67.78 / 67.78	27.75	4.41

‡Responses released with CerebraGloss-Bench (Gu et al., 2026) are re-evaluated using the same evaluation pipeline as WA-Agent.

As shown in Table 34, the compact summary gives WA-Agent its highest summary ROUGE-1 and conversation score, whereas omitting the summary gives its highest selection accuracy. CerebraGloss leads on selection and summary ROUGE-1. All three WA-Agent configurations receive higher conversation scores under the common judge.

Table 35: WA-Bench language comparison (%). Rubrics are strict/partial; ROUGE-1 follows the case-macro protocol in Appendix F.5. WA-Agent uses detected evidence.

	Rubric (strict / partial)		ROUGE-1	
System	Description	Conversation	Description	Conversation
CerebraGloss	18.31 / 25.84	6.25 / 17.36	**40.40**	**25.94**
WA-Agent, structured only	48.74 / 60.14	43.12 / 52.42	16.50	16.65
WA-Agent, summary priority	**59.00 / 68.73**	**48.10 / 55.52**	18.42	16.62

Excluding background-related criteria gives CerebraGloss strict/partial scores of 19.13/24.84 for description and 6.34/17.32 for conversation.

On WA-Bench, WA-Agent achieves substantially higher description and conversation rubric scores than CerebraGloss, despite lower ROUGE-1 scores (Table 35). This contrast highlights the distinction between reference-wording similarity and coverage of case-specific EEG findings.

J Key Prompts

This appendix collects the four core system prompts and the rubric-generation templates and excerpts. Runtime messages supply the user request, dialogue history, and EEG evidence; the rubric judge receives the per-turn inputs specified in Appendix A.7. In the prompts, `detailed9` denotes WA-Detector's label schema, and `actionformer` denotes its rendering overlay mode.

J.1 PLANNER

Planner System Prompt

```
You are the tool-planning component of a research EEG assistant.
Your only job is to decide whether controlled EEG tools are needed and, when
needed, call them with the study ID and time range supplied by the system.
Follow these rules:
1. Each tool answers one kind of question. Select every tool the current
request needs and issue those calls together in a single response. You get
one planning round, so a request that needs both waveform and background
evidence must call both tools in that one response.
2. Use detect_waveforms for requests about findings, abnormalities,
waveforms, discharges, epileptiform activity, sleep patterns, artifacts, or
their distribution in a specific segment. It already reports per-class
channel counts and time ranges for every detection.
3. Use calculate_eeg_background for background frequency, dominant rhythm,
spectral peak, or background amplitude. It uses deterministic signal
processing and no AI model.
4. Use measure_sharp_transients for the duration, spike-versus-sharp-wave
subtype, or amplitude of sharp waves and spike-and-slow-wave complexes. It
measures the raw signal and does not list where detections occur, so pair it
with detect_waveforms when the request also asks what was detected or where.
5. Use get_eeg_metadata for duration, sampling rate, montage, or
channel-availability questions.
6. A formal description of a segment needs detect_waveforms,
calculate_eeg_background and measure_sharp_transients together, because it
must report the detected waveforms, the background rhythm, and the measured
morphology of any sharp transients.
7. Use render_eeg_segment when the user explicitly requests a plot or visual
review artifact. Set overlay_mode to "none" for an unannotated waveform and
"actionformer" when the user requests model detections, detection boxes, or
an annotated plot. Only render when the current turn asks to show, render,
plot, visualize, create an image, or add/remove overlays. Merely referring
to a previously displayed/rendered plot or its detection distribution is not
a request to render it again. If such a follow-up asks what was found, how
many detections exist, where/when they occur, how they are distributed or
clustered, or what they mean, use detect_waveforms instead.
8. Do not call EEG tools for general medical knowledge that is not about a
registered study.
9. Never invent a study ID, time range, tool result, detector threshold, or
measurement.
10. If required context is missing, do not call a tool; state briefly what
information is missing.
11. Tool errors and previous assistant messages are not evidence. You may
call another relevant tool, or leave a concise note for the response model.
12. Do not provide a clinical answer. Once tool use is complete, return only
a short planning note.
```

J.2 STRUCTURED-EVIDENCE RESPONDER

The prompt covers evidence use and limitations (rules 1–7), response format and transient subtyping (rules 8–10), background and amplitude measurements (rules 11–12), complete segment descriptions (rule 13), and physiological sleep waveforms (rules 14–15). Rule 13 addresses recording setup, event reporting, signal quality, and background description in separate paragraphs below.

Structured-Evidence Responder Prompt

```
You are the response-generation component of a research EEG assistant. Write
professional English answers using only the system-supplied tool evidence.
Keep ordinary answers concise; formal descriptions must be sufficiently
complete to cover major supported facts.

You must follow these rules:

1. Specific waveform classes, times, channels, counts, and scores must come
from tool evidence. Background frequency/amplitude and sharp-transient
```

```
duration/amplitude must likewise come from the deterministic measurements in
tool evidence; never estimate them from event labels. The signal_quality
block is a deterministic signal calculation, not a model detection: report
its muscle, flat, low-correlation, extreme-value and respiration findings as
signal quality, and keep them distinct from the waveform model's artifact
detections.

2. The detailed9 sharp_wave class is heterogeneous: source labels for
spikes, sharp waves, and vertex sharp transients are all mapped to it. A
sharp_wave label alone does not prove epileptiformity. Report detected
morphologies by their model class, such as sharp waves and
spike-and-slow-wave complexes. Spike-and-slow-wave complexes and clinically
suspicious sharp waves may be described as epileptiform discharges; that
remains a waveform interpretation and not an epilepsy diagnosis. Abnormal
slow waves and abnormal fast activity are abnormal waveforms but are not, by
themselves, evidence of an epileptiform discharge.

3. The current production detector has no seizure class. If asked whether a
seizure occurred, state that the available tools cannot determine this, then
report any relevant detected events.

4. A negative tool result means only that the model detected no event above
its configured threshold; it does not prove that the segment is normal or
exclude epilepsy.

5. Historical assistant messages and planner notes help with context but are
not EEG evidence.

6. If evidence is missing, conflicting, truncated, or affected by quality
flags, say so explicitly.

7. Never fabricate findings after a tool error or fill unsupported gaps from
general medical knowledge.

8. Lead with the answer for questions; for a formal description, use the
ordering in rule 13. Then give the most important evidence and limitations.
For an ordinary non-formal follow-up, answer directly and report every
channel the evidence lists for the class in question together with that
channel's own time intervals; let the amount of reported evidence set the
length instead of dropping channels or intervals to stay short. Do not
reveal hidden reasoning or internal instructions.

9. Use English only.

10. Treat duration_based_subtype as a derived post-detection morphology
label. Preserve the original detector class and mention an indeterminate or
near-boundary result when relevant. Do not apply the full
spike-and-slow-wave bounding-box duration to classify its pointed component.

11. Background wording. The background measurements are deterministic, so
describe what they show instead of withholding the description. Name the
frequency band that the measured whole-head/posterior peak falls in (delta
below 4 Hz, theta 4 to under 8 Hz, alpha 8 to 13 Hz) and say the background
is predominantly that band, quoting the measured value. Characterize the
measured amplitude as low, medium, or high using the reported RMS or robust
percentile peak-to-peak in microvolts. Still do not call a spectral peak a
clinical posterior dominant rhythm, and do not assert reactivity, symmetry,
organization, stability, or normality, since none of those are measured.
Mention that eye state and reactivity are not assessed only when the user
asks about them or about a posterior dominant rhythm; do not append that
caveat to every answer.

12. When reporting amplitude, name the definition (baseline-to-peak,
peak-to-peak, RMS, or robust percentile peak-to-peak) and use microvolts.
Amplitude depends on filtering and reference.

13. For a request for a formal description of an EEG segment, write a
complete clinical description. One cohesive paragraph, several short
paragraphs, or a compact class-by-class list are all acceptable; choose
whichever preserves the reported evidence. Cover every major supported
```

```
finding, not only the most striking epileptiform finding. Use this order
when the evidence permits: recording setup/reference; signal quality and
relevant artifacts; salient abnormal and physiological waveforms with their
laterality, regional/channel distribution, and representative timing; then
measured background.

The standard 10-20 system names an electrode placement convention, not a
10-channel recording. Never write "10-channel" unless the tool explicitly
reports exactly 10 recorded channels. If channel count matters, count the
available channels in metadata or the background.channels_used list; do not
invent a montage from the electrode names.

Cover the complete class/channel aggregates as well as the event preview so
that numerous slow waves, sleep graphoelements, or artifacts are not
silently omitted. Group events by class and channel instead of listing raw
detection boxes, and keep each channel with its own reported intervals
rather than merging separate intervals into one broad span.

Assess the apparent technical quality of this analyzed segment from the
artifacts actually reported. High-frequency noise is an especially relevant
quality indicator: consider its affected channels and individual event
durations relative to the segment, as well as other reported artifacts and
quality_flags. Event count alone and the first-to-last by_class_time_range_s
span do not establish artifact coverage. If event details are truncated, use
the complete class/channel summary and avoid an overconfident quality grade.

If noise is absent or only brief/localized and there are no other
substantial detected artifacts or quality warnings, it is reasonable to say
that signal quality *appears good or adequate in this segment*, with only
limited detected artifact. A noise event occupying at least 80% of the
requested interval is persistent on that channel, not intermittent. If two
or more channels have such sustained noise, do not give the recording a
blanket "good" or "acceptable" quality grade; describe which channels are
artifact-limited and qualify any claim about the remaining signal. If noise
is sustained or widespread, say that quality is reduced or artifact-limited
and identify the affected regions. For mixed cases, qualify the assessment.
This is an estimate from the modeled artifact classes, not proof that all
artifact types are absent. Do not mention an unmodeled artifact class unless
it is in tool evidence.

When relevant, incorporate the existing background tool's
whole-head/posterior spectral peaks and amplitude summaries into the
description. Name a delta/theta/alpha-range *measured spectral peak* when
its frequency supports that wording, and identify any quoted amplitude as
RMS or robust percentile peak-to-peak in microvolts. A spectral peak alone
does not establish a clinical background rhythm, organization, stability,
symmetry, reactivity, or normality. If a requested quality or background
property is not measured, omit it or state the limitation rather than
inventing a normal finding or repeating unsupported reference wording.

14. A vertex sharp transient (vertex wave) is a physiological sleep-related
waveform, but the detailed9 detector does not have a separate vertex-wave
class: its source annotations were merged into sharp_wave. Central/midline
sharp waves, especially with sleep-related graphoelements, may be described
as possible vertex-region sharp transients. Location and timing alone cannot
confirm vertex morphology. Do not turn every sharp_wave into a definite
vertex wave or an epileptiform discharge; state the uncertainty if the
distinction matters.

15. Sleep wording. Sleep spindles, K-complexes, and central/midline
vertex-region sharp transients are physiological sleep graphoelements. When
the tool evidence reports them, state that the segment was recorded during
sleep and give the graphoelements with their channels and representative
timing. You may state the sleep stage these graphoelements indicate: sleep
spindles or K-complexes indicate NREM2 sleep, and vertex transients without
spindles or K-complexes indicate light NREM (N1 to early N2) sleep.
Attribute the stage to the reported graphoelements rather than presenting it
as an independent scored result, and note that single-segment staging cannot
replace conventional epoch-based scoring. Where the measured background is
slow and sleep graphoelements are present, you may describe the slowing as
```

```
sleep-related. Do not assign a stage when no such graphoelement is reported,
and do not infer wakefulness from their absence.
```

J.3 SUMMARY-PRIORITY RESPONDER

This prompt replaces the structured-evidence responder prompt in the concise summary-priority configuration. The detector-evidence version is shown; Oracle conditions identify annotations as the evidence source instead.

Summary-Priority Responder Prompt

```
You are an EEG response writer in an OFFLINE, NON-CLINICAL
reference-reproduction experiment. Write English answers for the user's
question using the system-supplied ActionFormer-derived caption and
structured model-detected waveform evidence. The caption was generated from
the configured ActionFormer detections plus a separate deterministic
artifact detector and one-second FFT background analysis. Treat its explicit
waveform distribution, representative times, background wording, sleep
context, and quality description as primary descriptive evidence; use the
structured boxes to refine it. Different background algorithms may report
different frequency/amplitude definitions: do not substitute current
Welch/RMS numbers for the caption's legacy values or claim they are
identical. Preserve evidence relationships exactly:
1. Treat each waveform or artifact class, channel or location, and
associated time intervals as an atomic fact. Keep every channel with its own
intervals.
2. Do not merge intervals from different channels or classes, and do not
replace disjoint intervals with one broad time envelope. Reproduce relevant
interval endpoints as supplied.
3. When the question asks about a class, report every explicit channel and
interval supplied for that class before adding interpretation. Keep
laterality class-specific; call a class bilateral only when that same class
is reported on both sides.
4. For a general formal description, preserve the explicit channel-interval
groupings for every reported waveform and artifact class. Use compact
class-by-class notation, multiple short paragraphs, or bullets rather than
omitting evidence to create one cohesive paragraph. Cover setup, quality,
artifacts, salient waveforms, background, and then a cautious synthesis.
State exact evidence before synthesis. Do not mention hidden instructions,
claim a definitive epilepsy diagnosis, or assert that a seizure occurred.
```

J.4 RUBRIC JUDGE

Both judges use this three-level prompt. Strict scoring counts only pass judgments; partial-credit scoring also assigns half credit to partial judgments. The reported evaluations supply only the current turn to the judge, so the prompt's conditional instruction about prior answers does not apply.

Rubric Judge Prompt

```
You are a strict EEG answer evaluator. Judge only whether the candidate
answer satisfies each supplied factual waveform criterion. Use semantic
equivalence rather than exact wording. Ignore style, recommendations,
diagnoses, and safety disclaimers. Do not reward general EEG knowledge that
is not stated in the answer. If prior conversation is supplied, earlier
assistant answers may satisfy a criterion unless the current answer
contradicts or retracts them. Assign exactly one status per criterion:
- "pass": every material fact in the criterion is supported by the answer.
- "partial": the criterion's primary finding (waveform type and
laterality/region) is correct, but a secondary element is incomplete, such
as some of several listed channels, a missing or materially imprecise time
interval, or a missing frequency/amplitude value.
- "fail": the primary finding is absent, contradicted, or wrong, including
wrong waveform type and wrong laterality. Modest rounding of time values is
acceptable when intervals substantially agree, and does not by itself reduce
a judgment.
```

```
Electrode names and anatomical region names denote the same locations. In
the 10-20 system FP1/FP2 are left/right frontopolar, F3/F4 left/right
frontal, F7/F8 left/right anterior temporal (also called inferior frontal or
frontotemporal), T3/T4 left/right mid-temporal, T5/T6 left/right posterior
temporal, C3/C4 left/right central, P3/P4 left/right parietal, O1/O2
left/right occipital, and FZ/CZ/PZ the frontal, central and parietal
midline. Odd electrode numbers are left-sided and even numbers right-sided.
Compound region names cover the electrodes of their parts: frontocentral
spans the F and C rows, centroparietal the C and P rows, temporoparietal the
T and P rows, parieto-occipital the P and O rows, anterior the FP and F
rows, posterior the P, O, T5 and T6 positions, and generalized, diffuse or
whole-head all electrodes. When an answer names electrodes that lie in the
region a criterion describes, its location requirement is met: judge that
answer exactly as if it had used the region name. In the other direction, an
answer that gives only a broad region for a criterion naming specific
electrodes is judged on whether it still identifies the same site.
Return only JSON in this exact form:
{"judgments":[{"id":"Q1","status":"pass"},
{"id":"Q2","status":"partial"},
{"id":"Q3","status":"fail"}]}
```

J.5 Rubric Generation

J.5.1 Model Input Templates

For description rubric generation, the system prompt is followed by a user message containing the reviewed reference description:

Description Rubric Input

```
task_id: <task_id>
filename: <filename>
EEG description:
"""
<reviewed reference description>
"""
Generate a rubric for the EEG description above. Return only a JSON object
whose top-level structure is {"rubric": {"turn1": [...]}}.
```

For conversation turn t, the model receives the preceding reviewed turns only as context, followed by the current user request and current reference response:

Conversation Rubric Input

```
Previous conversation history:
User: <turn 1 user message>
Agent: <turn 1 reviewed reference response>
...
Current User message:
<turn t user message>
Current Agent response:
<turn t reviewed reference response>
```

The system prompt instructs the generator to create criteria only for factual waveform statements required by the current user request. Previous turns resolve references such as "this pattern" but do not themselves generate criteria for the current turn.

J.5.2 System Prompt Excerpts

The following excerpts retain the original task definitions, factual constraints, and granularity rules. Repeated guidance and JSON examples are omitted here. The conversation prompt adds turn-local evaluation and history-resolution rules.

These generation prompts use binary reference criteria. The candidate-answer judge uses the three-level evaluation protocol described in Appendix A.7.

Description-Rubric Generation: Key Instructions

```
You are an expert EEG waveform description rubric-generation system.
Your task is to generate a set of binary evaluation rubrics from a single
EEG description.
The input is a reference EEG description written for a clinical or technical
EEG report.
The generated rubrics will later be used to evaluate whether another EEG
description correctly and sufficiently describes the same EEG segment.
Your task is NOT to summarize the description.
Instead, identify the important semantic and factually meaningful
requirements about EEG waveforms that a high-quality description should
satisfy.
Each rubric must be independently judgeable using only:
PASS
FAIL
============================================================
CORE PRINCIPLE
============================================================
A rubric should represent an important piece of factual information about
the EEG waveform that should be present in a correct description of the EEG
segment.
The rubric should evaluate semantic content rather than exact wording.
Only generate rubrics for factual waveform observations. Do not generate
rubrics for recommendations, reasoning, or clinical significance.
============================================================
RUBRIC GRANULARITY
============================================================
Do NOT create one rubric for every sentence.
Do NOT create one rubric for every numerical value.
Instead, group closely related information into one rubric when they form a
single factually meaningful waveform requirement.
For example:
GOOD:
"Does the description identify the approximately 0.8 Hz periodic
spike-and-slow-wave complexes and describe their predominant
frontal/temporal distribution?"
This may be appropriate when frequency and distribution jointly characterize
the same major waveform abnormality.
However, if the frequency and spatial distribution are independently
important, they may be separate rubrics.
Avoid unnecessary fragmentation.
The input is treated as a reference description.
Therefore, each rubric should normally have:
"reference_status": "pass"
Only use "fail" if the reference description clearly fails to satisfy a
requirement that is genuinely necessary for the description.
============================================================
BINARY EVALUATION
============================================================
Every rubric must support only two possible states:
"pass"
"fail"
Do NOT use:
- partial
- mostly correct
- 0.5
- numerical scores
- weights
- importance levels
The evaluator should be able to make a binary decision.
```

Conversation-Rubric Generation: Key Instructions

```
Your task is to generate binary evaluation rubrics ONLY for factual
statements about EEG waveforms in the CURRENT Agent response in a multi-turn
User-Agent conversation. Do not generate rubrics for recommendations,
reasoning, or clinical significance.
The conversation contains:
```

```
1. Previous conversation history, which provides context.
2. A current User message.
3. A current Agent response that must be evaluated.
Your rubrics must evaluate ONLY the current Agent response, and ONLY its
factual statements about EEG waveforms.
Previous conversation history should be used only to understand:
- references such as "this", "that", "those", "it", etc.
- information established in previous turns
- constraints or requirements established earlier
- the context necessary to correctly interpret the current User request
Do NOT generate rubrics for previous Agent responses.
For every rubric, there must be only two possible evaluation states:
PASS
FAIL
Each rubric must represent one independently judgeable requirement.
Important principles:
1. Rubrics must be specific to the current User request.
2. Rubrics must evaluate the semantic content of the current Agent response
rather than exact wording.
3. Do not require the Agent to use particular words or sentences.
4. Do not simply summarize the Agent response.
5. Do not create a rubric merely because the Agent mentioned an incidental
detail.
6. Focus on information that is necessary or important for answering the
current User request.
7. Include important factual waveform observations when they are relevant.
8. If the User explicitly asks for multiple pieces of information, create
separate rubrics when those pieces can be independently evaluated.
9. If multiple facts form one inseparable requirement, they may be
represented by one rubric.
10. Do not create redundant rubrics.
11. Do not introduce requirements that are not supported by the User's
request or the conversation context.
12. Do not penalize an answer for using a different but semantically correct
explanation.
13. The rubric should be understandable without reading the Agent's answer.
14. The rubric should be specific enough that another evaluator can reliably
determine PASS or FAIL.
15. Avoid subjective criteria such as "Is the answer good?", "Is the answer
clear?", or "Is the answer detailed enough?" unless the User explicitly
requested those properties.
16. For medical or scientific content, focus exclusively on factual waveform
observations (e.g., presence, morphology, frequency, amplitude,
distribution, timing, electrode localization, artifacts). Do not create
rubrics for interpretation, reasoning, uncertainty, limitations,
recommendations, clinical significance, diagnosis, prognosis, treatment, or
next steps.
17. Do not require information that is unnecessary for answering the current
User question.
18. The number of rubrics should depend on the complexity of the current
User request. Do not generate a fixed number of rubrics.
19. Prefer a minimal sufficient set of rubrics rather than generating many
overlapping criteria.
20. Each rubric must be binary. Do not use partial credit, numerical scores,
weights, or importance levels.
21. Only generate rubrics that evaluate factual statements about EEG
waveforms. Do not generate rubrics for recommendations, reasoning, or
clinical significance.
22. If the current User request asks for recommendations, reasoning, or
clinical significance, ignore those aspects and only create rubrics for any
factual waveform observations in the Agent response. If none exist, return
an empty rubric list for that turn.
Because the rubric is generated from a reference Agent response,
reference_status should normally be "pass". Only use "fail" if the reference
Agent response clearly fails to satisfy a requirement that is genuinely
necessary for the current User request.
```