

Figure 7: Activated Channels in Each Brain Area

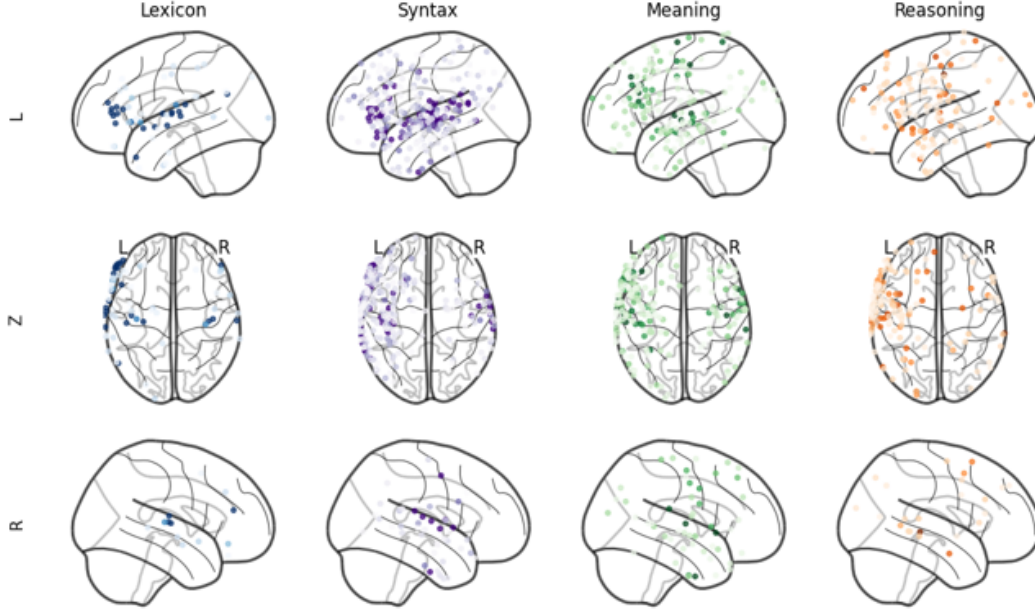


Figure 8: **Brain maps showing the spatial distribution of electrodes dominated by each linguistic feature, across three anatomical views.** Electrodes are colored according to their most selective feature based on peak z-scored encoding correlation values, computed after removing word rate confounds. Each feature (Lexicon, Syntax, Meaning, Reasoning) is shown in a separate column, and each row corresponds to a different cortical view: left lateral, dorsal (axial), and right lateral. Only electrodes exceeding a z-threshold of 2.58 are shown.

When assigning each electrode to the linguistic feature for which it shows the highest encoding z-score, we observe clear regional differentiation across features. Electrodes dominated by shallow lexical and syntactic features (Lexicon, Syntax) are primarily localized to canonical perisylvian language areas, including the superior temporal gyrus (STG) and the inferior frontal gyrus (IFG), particularly its ventral portion. Semantic feature-selective electrodes (Meaning) extend more dorsally within IFG and are distributed more broadly across the temporal and frontal cortices, consistent with prior accounts of distributed semantic representation. In contrast, reasoning-related electrodes (Reasoning) engage distinct regions, notably including the superior frontal gyrus (SFG) and occipital areas, suggesting recruitment of domain-general and potentially visual-associative mechanisms unique to high-level inferential processes.

B Compute Resources and Execution Time

We report here the compute infrastructure and approximate runtime for each stage of our pipeline to support reproducibility.

- **Hidden State Extraction:** Performed using $4 \times$ NVIDIA L40 GPUs. Extracting hidden states for around 164,000 tokens from the Qwen2.5-14B model. The extraction took approximately 40 minutes and required around 4×30 GB of GPU memory.
- **Layer-wise Probing:** Conducted using $1 \times$ NVIDIA L40 GPUs. Each BLiMP task paradigm took 2.2 minutes on average, totaling around 57.2 minutes for the selected 26 paradigms focusing on syntax. COMPS-BASE consisting of 49,340 English sentence pairs took 1 hour and 47 minutes on average. COMPS-WUGS-DIST consisting of 27,792 pairs took 1 hour on average.
- **Residual Embedding Construction:** Ridge regression training was done using Scikit-learn on CPU with 30 GB of memory. Each regression model took less than 10 minutes to