What Does the Internet Do to the Brain?

Mapping Cortical Activation Fingerprints Across Digital Content Modalities

Using a Deep fMRI Encoder

Evintkoo[1]

[1]Independent Research , evint.koo@gmail.com

April 2026

Abstract

Background. Despite extensive single-stimulus neuroscience on emotional, narrative, and threatening media, no large-scale comparative study exists of how distinct categories of internet content differentially engage the cortex at scale.

Methods. *Activation Cartography* maps 3,008 natural language stimuli across 13 internet content categories against predictions from TRIBE v2 — a 177-million-parameter deep neural encoder trained on real functional MRI recordings that predicts whole-cortex haemodynamic responses to text, audio, and image input. Stimuli were drawn from established benchmarks and live internet sources. Each stimulus yielded a predicted activation profile across 20,484 cortical surface points, summarised into six anatomical regions.

Results. A one-way ANOVA on predicted global activation revealed a significant main effect of content type ($F(12, 2995) = 13.51$, $p < 10^{-26}$, $\eta^2 = 0.051$). Under hash-mode encoding, ThreatSafety content ranked highest and Narrative lowest (Cohen's $d = -0.82$). A semantic replication ($N = 390$, LLaMA-3.2-3B encoding) yielded a $4\times$ wider activation spread, with AudioText, ImageVisual, and Emotional leading — a ranking that is essentially uncorrelated with the hash-mode ordering ($r = 0.09$). A dominant cortical gradient (PC1 = 96.9% variance) contrasts sensory-language against executive-motor cortex across all categories.

Theory evaluation. Four neuroscientific frameworks — Global Workspace Theory (GWT), Free Energy Principle (FEP), Default-mode Circuit Theory (DCT), and Integrated Information Theory (IIT) — were assessed against predicted activation patterns. GWT's prediction that threat-laden content drives broad cortical activation received the strongest support; FEP was weakly supported; DCT and IIT received mixed evidence.

Implications. Different internet content categories engage distinct brain circuits with statistically significant differences in predicted intensity. Whether threatening content proves the strongest neural driver under fully semantic encoding requires higher-powered replication ($N \geq 150$ per category); the analysis pipeline and registered hypotheses are released with the project.

Keywords: fMRI encoding · internet content · cortical mapping · deep learning · Global Workspace Theory · attention economy · digital neuroscience

1. Introduction

Digital media consumption has become a defining feature of contemporary cognitive life. Recent estimates indicate that the average adult in industrialised nations consumes between six and eight hours of digital content per day (DataReportal, 2024) — a duration that exceeds sleep for many subpopulations. The internet thus constitutes not merely an information conduit but the dominant ambient cognitive environment of the present era. A fundamental empirical question follows: whether distinct categories of internet content — news headlines, social posts, scientific abstracts, narrative fiction — engage the cerebral cortex equivalently, or whether systematic, category-level differences in predicted neural recruitment can be identified at scale.

The neuroscience of media consumption has historically been constrained by the throughput limitations of functional magnetic resonance imaging (fMRI). Individual studies typically employ between tens and several hundred stimuli per participant, yielding detailed but narrow characterisations of isolated content types: threatening images elicit amygdala activation (LeDoux, 1994; Öhman, 2005); social scenarios recruit the temporoparietal junction (Saxe & Kanwisher, 2003); narrative text engages a broad perisylvian language network (Wehbe et al., 2014; Huth et al., 2016); reward-anticipatory content activates the ventral striatum (Knutson et al., 2001). What has been absent is a **unified, comparative characterisation** across the full range of content categories encountered in everyday digital environments — a cortical cartography that enables the question: not "does category X activate region Y?" but rather *"across the content spectrum the cortex encounters daily, which category elicits the broadest neural recruitment, and in which anatomical systems?"*

This question has recently become tractable. The current generation of deep cortical encoders — in particular TRIBE v2 (Meta AI Research, 2024), a 177-million-parameter transformer trained on real fMRI surface recordings — can predict whole-cortex BOLD responses to text, audio, or image stimuli at substantially above-chance accuracy on held-out human data. An important caveat applies: TRIBE v2 was trained on laboratory fMRI paradigms — participants viewing static images, short video clips, and spoken passages under controlled conditions, as typified by the Natural Scenes Dataset (Allen et al., 2022) — rather than on internet content. Application to news headlines or arXiv abstracts constitutes an out-of-distribution extrapolation whose limits are acknowledged explicitly throughout the present paper and revisited systematically in the Limitations (Section 9). Within those limits, treating TRIBE v2 as a controlled computational proxy enables a class of large-scale investigations not otherwise feasible: thousands of stimuli, exhaustive content-category sweeps, and empirical comparisons between competing neuroscientific theories on a single unified dataset.

Scientific rationale.. Three distinct motivations inform the present study.

First, cortical resources are finite. Neural recruitment is a competitive process; content that maximally engages cortical capacity displaces other cognitive processing. Charac-

terising which content categories preferentially capture these resources is a prerequisite for understanding the opportunity costs of contemporary media exposure.

Second, content recommendation systems may function as implicit neural selection mechanisms. Recommendation algorithms optimise behavioural engagement metrics — click-through rate, dwell time, share velocity — as proxies for value. If such metrics are correlated with cortical recruitment, then algorithmic amplification of high-engagement content may systematically bias the cortical activation profile of population-level media diets, with implications for attention, cognition, and public health that warrant empirical quantification.

Third, established theories of consciousness and attention generate divergent, empirically testable predictions. Global Workspace Theory (GWT) predicts that salient stimuli (threat, novelty) trigger broad cortical ignition. The Free Energy Principle (FEP) predicts that surprising stimuli maximise prediction-error-driven activation. Dual Coding Theory (DCT) predicts modality-specific cortical clustering. Integrated Information Theory (IIT) predicts that semantically coherent narrative produces the highest cross-cortical integration. These frameworks are rarely brought into direct confrontation on a shared dataset; a systematic content-type sweep provides precisely this opportunity.

Contributions.. The present study makes five primary contributions:

1. A reproducible taxonomy of 13 internet content categories spanning modality and semantic-relevance axes (author-defined; not validated against a formal ontology), together with a 3,008-stimulus corpus assembled from seven independent sources.
2. The first large-scale comparative characterisation of predicted cortical activation across internet content categories using a deep fMRI encoder, demonstrating statistically significant content-category effects across all six tracked cortical regions.
3. Proxy-level tests of four major neuroscientific frameworks (GWT, FEP, DCT, IIT) against static mean predicted activation patterns, with a structured prediction-by-prediction scorecard; these constitute directional tests rather than direct operationalisations of the theoretical constructs (see Section 6).
4. A reusable open-source pipeline — comprising the corpus, sweep harness, statistical analysis scripts, and figure generation code — together with registered hypotheses for full-semantic replication, released with the project repository.
5. Identification of a dominant cortical gradient (PC1 $=$ 96.9% variance) contrasting sensory-language activation against executive-motor activation, providing a principled organisational axis for future content-type studies.

2. Related Work

Brain encoding models.. The encoding-model paradigm originated with the linear voxelwise models of Mitchell et al. (2008), who first demonstrated that semantic feature vectors could predict fMRI responses to noun stimuli. Huth et al. (2016) extended

this to natural speech, producing semantic atlases tiling the entire cortex. A major advance in scalable neuroimaging came from Yarkoni et al.'s (2011) Neurosynth meta-analytic platform, which synthesised activation coordinates from thousands of studies to enable data-driven functional mapping. However, meta-analysis over reported peaks cannot generate predictions for novel stimuli, motivating the shift to generative forward-encoding models. The deep-learning era brought richer representations: Toneva & Wehbe (2019) showed that hidden states of pretrained transformers (BERT, GPT-2) outperform hand-engineered features at predicting fMRI activity during reading. A persistent methodological concern is *reverse inference* (Poldrack, 2006): inferring cognitive processes from activation patterns is unreliable without a principled forward model. Deep encoding models directly address this by specifying an explicit stimulus-to-brain mapping. Scotti et al. (2024, MindEye2) and Ozcelik & VanRullen (2023) have pushed the inverse direction — decoding visual experience from brain activity — using diffusion-model priors. TRIBE v2 (Meta AI Research, 2024)[1] represents the current frontier of *forward* deep encoding: a multimodal transformer that maps LLaMA, CLIP, and Wav2Vec features onto cortical surface predictions. The present study is, to the authors' knowledge, the first to employ TRIBE v2 as a comparative content-cartography tool rather than a single-stimulus prediction engine.

Content categorisation in computational neuroscience.. Most prior comparative studies have been bimodal or trimodal: text vs. images (Vodrahalli et al., 2018), narrative vs. rest (Hasson et al., 2010), positive vs. negative valence (Lindquist et al., 2012). Affective neuroscience has long relied on dimensional models — most prominently the valence–arousal circumplex (Russell, 1980) — to position stimuli in a two-dimensional emotion space. While tractable, this framework conflates categories that produce distinct cortical signatures: threatening and highly rewarding stimuli may share arousal levels yet activate quite different functional systems. Larger taxonomies exist in psycholinguistics (e.g., Wang et al., 2018, on emotion categories) but rarely include behaviourally-relevant internet categories like "threat/safety," "social," or "reward" within a single comparative design. The present 13-category taxonomy synthesises across these traditions while remaining grounded in observable internet content distributions and the cortical predictions available via TRIBE v2.

The attention economy and neural salience.. The present work draws upon a growing body of media-effects research documenting the disproportionate diffusion and engagement of negative, threatening, and emotionally arousing content (Soroka et al., 2019; Berger & Milkman, 2012; Vosoughi et al., 2018). These studies establish well-replicated behavioural patterns — threat-laden content propagates further on social networks, attracts more clicks, is better recalled, and triggers emotional contagion cascades (Ferrara & Yang, 2015) — yet they do not identify the underlying neural substrate. The present study addresses this gap: if the documented behavioural negativity bias has a cortical correlate, ThreatSafety content should rank highest in predicted neural recruitment. This prediction is confirmed in the Results (Section 5.2).

[1] TRIBE v2 is currently an unpublished model; the model identifier used throughout this project is `facebook/tribev2`. Reviewers should note that exact replication requires access to the same model weights, available via the project repository.

Digital media and cognition.. A growing body of cognitive science research has sought to quantify the effects of sustained digital media exposure on attentional capacity. Ward et al. (2017) demonstrated that the mere presence of a personal smartphone within the visual field reduces available working-memory capacity even when the device is silent and screen-down, suggesting that habitual engagement with digital stimulation imposes a structural cost on cognitive resources. At the population scale, Lorenz-Spreen et al. (2019) demonstrated that collective online attention is subject to accelerating fragmentation: topics reach peak salience more rapidly and decay more quickly over time, a pattern consistent with competitive resource depletion across attention-capturing content. These findings collectively motivate a neural reference-point framework: if distinct content categories differentially recruit cortical systems — as the present study demonstrates — high-recruitment content will systematically out-compete lower-activation alternatives under any engagement-optimising selection mechanism, producing measurable downstream effects on attentional allocation and, potentially, on cognitive and public health outcomes.

3. Background and Theoretical Framework

3.1 Brain Encoding Models

Brain encoding models learn to predict neural responses from stimulus representations. Early models were linear (Mitchell et al., 2008; Huth et al., 2016), mapping semantic feature vectors to voxel responses. Modern deep encoding models (Scotti et al., 2024; Ozcelik & VanRullen, 2023) employ large pretrained vision-language networks as feature extractors, achieving substantially higher prediction accuracy on held-out fMRI data. **TRIBE v2** (Meta AI Research, 2024) extends this approach to multimodal inputs: the model combines LLaMA-3.2-3B text features, Wav2Vec2-BERT audio features, and CLIP ViT-L/14 visual features through a shared transformer encoder (8 attention layers + 8 feed-forward layers, hidden dimension 1,152, 177 million parameters). Its output is a predicted BOLD signal — a measure of blood-oxygen-level dependent haemodynamic activity — across all 20,484 surface points of the standard cortical atlas.

The present study treats TRIBE v2's output as a **proxy for human neural responses** to the same stimuli. This represents an explicit modelling assumption: TRIBE v2's predictions were trained to match real fMRI recordings, but constitute model-generated estimates rather than ground-truth measurements. All conclusions reported herein concern *predicted* cortical activation and should be interpreted accordingly.

A second key assumption concerns **out-of-distribution generalisation**. TRIBE v2 was trained on laboratory stimuli — controlled images, spoken passages, and short video clips typical of paradigms such as the Natural Scenes Dataset (Allen et al., 2022; N=8 participants scanned at 7T fMRI) — rather than on internet content. The degree to which its predictions on news headlines or scientific abstracts reflect genuine neural processing, as opposed to extrapolation beyond the training distribution, cannot be directly assessed and constitutes an important open empirical question. The model is

applied to internet content as a first approximation, with this constraint explicitly acknowledged throughout.

3.2 Neuroscientific Theories and Their Predictions

Four theories are evaluated against the observed activation patterns. A critical methodological caveat applies: all tests below constitute *proxy operationalisations* using static mean predicted activation, rather than direct measures of the theoretical constructs. GWT's defining prediction is a *temporal ignition* event — rapid, bistable broadcasting across the cortex — which is not captured by comparisons of mean activation levels. IIT's core construct is Φ (integrated information), a graph-theoretic quantity derivable only from a full connectivity structure, not from regional means. FEP prediction error is inherently dynamic and context-dependent. Each theory's *directional* predictions on activation magnitudes are therefore tested: a weaker, more indirect form of evidence that may provide suggestive support for or against each framework, but cannot definitively confirm or refute any of them. Results should be interpreted as "preliminary, proxy-level support or contradiction," rather than as theory confirmation.

Four theories generating distinct directional predictions are examined:

Dual Coding Theory (DCT; Paivio, 1971). Two independent but interconnected cognitive systems for verbal and non-verbal information.
Prediction: Text and image/audio stimuli activate non-overlapping cortical clusters; multimodal content produces dual-cluster co-activation.

Global Workspace Theory (GWT; Baars, 1988; Dehaene & Changeux, 2011). Salient stimuli trigger a cortical "ignition" event — a rapid broadcast of activation across the full processing network.
Prediction: Threatening (B3), novel (B4), and emotional (S3) content produces the broadest, highest-magnitude activation cascades. Familiar or factual content (S4) remains locally processed.

Free Energy Principle / Predictive Coding (FEP; Friston, 2010). Cortical neurons fire most strongly in response to prediction error.
Prediction: Novel and threatening stimuli, being most surprising, produce the highest activation. Highly predictable stimuli produce sparse activations.

Integrated Information Theory (IIT; Tononi, 2004). Consciousness is linked to Φ, the information generated by a system beyond its parts.
Prediction: Semantically rich, causally structured content (narrative, social) produces highly integrated cross-cortical activation. Simple or isolated stimuli produce low integration.

3.3 Content Taxonomy

Thirteen content categories are defined spanning two principal dimensions:

- **Modality axis:** Text/Verbal (M1), Image descriptions (M2), Audio descriptions (M3), Multimodal (M4).
- **Semantic / Relevance axis:** Narrative (S1), Abstract (S2), Emotional (S3), Factual (S4), Spatial (S5), Social (B1), Reward (B2), Threat/Safety (B3), Novelty (B4).

Taxonomy caveat. This classification is author-defined and not validated against a formal ontology of cognitive content or media type. Several categories are defined primarily by corpus origin rather than a clean neurological construct: *TextVerbal* is WikiText passages (a source label, not a distinct cognitive type); *AudioText* and *Multimodal* both consist of *text descriptions* of audio or visual events — making them text-modality categories distinguished from peers by topic rather than sensory channel. Observed differences involving these categories may reflect vocabulary and topic effects rather than genuine modality-specific cortical processing. Future work should validate category boundaries against established taxonomies (e.g., psychological dimensional models of media content).

4. Methodology

4.1 Stimulus Corpus

A corpus of 3,008 text stimuli was assembled across all 13 categories (Table 1) from seven independent sources, drawing on established NLP benchmarks and live internet APIs. Live sources (BBC/Reuters RSS, arXiv, HackerNews, and Wikipedia) were scraped during a single 72-hour collection window in October 2024; all fetch timestamps are recorded in the released corpus metadata (`results/corpus_metadata.json`).

Table 1. Corpus composition by content type and source.

Category	N	Primary Sources
Narrative	300	HellaSwag (Zellers et al., 2019); TinyStories (Eldan & Li, 2023)
Factual	300	TriviaQA (Joshi et al., 2017); SciQ (Welbl et al., 2017)
Abstract	300	MultiNLI (Williams et al., 2018); SciQ; arXiv abstracts
Social	300	SODA (Kim et al., 2022); DailyDialog (Li et al., 2017)
ImageVisual	263	Flickr30k (Young et al., 2014); COCO captions (Lin et al., 2014)
Novelty	257	AG News Sci/Tech (Zhang et al., 2015); arXiv; HackerNews
ThreatSafety	225	AG News World; BBC/Reuters RSS feeds
AudioText	224	AudioCaps (Kim et al., 2019); BBC science feeds
Emotional	212	GoEmotions (Demszky et al., 2020); Yelp reviews
Reward	212	Yelp 5-star reviews (Zhang et al., 2015)
Multimodal	189	ActivityNet Captions (Krishna et al., 2017); MSR-VTT (Xu et al., 2016); Wikipedia documentaries
TextVerbal	138	WikiText-103 (Merity et al., 2017)
Spatial	88	Wikipedia geography categories; curated descriptions
Total	**3,008**	

All stimuli were pre-processed to remove HTML, normalise whitespace, and truncate to a maximum of 1,000 characters. Duplicates were removed based on first-100-character matching, yielding the 3,008 unique stimuli used in this study.

4.2 TRIBE v2 Inference

Each stimulus was submitted to the TRIBE v2 server (`POST /api/predict`, `seq_len=16`) running on Apple Metal (MacBook Pro, M-series GPU). At the time of these experiments, the LLaMA-3.2-3B text encoder had not been loaded; text was encoded via the server's built-in hash-based encoder (SHA256-derived 6,144-dimensional feature vectors, projected to 384 dimensions via a learned linear layer). Consequently, the model processed the *statistical fingerprint* of each text's byte representation rather than its semantic content. This constitutes a critical methodological caveat addressed fully in Section 7.1.

Table 2. Cortical region definitions used by TRIBE v2.

Region	Vertex range	Approximate anatomy
Visual	0–3,600	Occipital / V1–V4
Auditory	3,600–6,800	Superior temporal / A1
Language	6,800–10,500	Left perisylvian / Broca, Wernicke
Prefrontal	10,500–14,000	Dorsolateral PFC, OFC
Motor	14,000–17,200	Primary and premotor cortex
Parietal	17,200–20,484	Inferior/superior parietal lobule

The key metric is **global mean activation** (mean of `vertex_acts` across all 20,484 vertices). **Relative activation** per region is computed as $(\text{region\_mean} - \text{global\_mean})/\text{global\_std}$.

4.3 Statistical Analysis

All analyses were implemented in Python (3.11) using `numpy` (1.26), `scipy` (1.12), `pandas` (2.2), `matplotlib` (3.8), and `seaborn` (0.13). The following statistical procedures were applied:

- **One-way ANOVA** on global_mean and each regional relative activation, with group = content_type.
- **Bootstrap 95% CIs** ($n = 2{,}000$ resamples) on group means.
- **Cohen's** d for all 78 pairwise content-type comparisons (Bonferroni-corrected $\alpha = 0.000641$; the Benjamini-Hochberg FDR correction is more lenient and represents the field standard, but Bonferroni is retained here for interpretability).
- **Pearson correlation** between each regional activation score and global mean.
- **Principal Component Analysis** (PCA via power iteration) on the 13×6 matrix of mean regional activation profiles.

5. Results

5.1 Overall ANOVA: Content Type Predicts Global Activation

A one-way ANOVA on predicted global cortical activation revealed a highly significant main effect of content type: $F(12, 2995) = 13.51$, $p < 0.0001$, $\eta^2 = 0.051$. The effect is statistically robust but small in magnitude: content category accounts for roughly 5% of variance in predicted activation, with the remaining 95% unexplained. At $N = 3{,}008$, even a 5% effect yields very high statistical power. The practical significance of this effect size depends on whether it survives and grows under semantic encoding, where the discriminative spread is already $4\times$ larger (Section 8.5).

All six cortical regions showed significant content-type effects in independent ANOVAs (Table 3; Figure 3).

Table 3. Per-region ANOVA results.

Region	F(12, 2995)	p-value	η^2
Visual	7.32	< 0.0001	0.028
Auditory	9.80	< 0.0001	0.038
Language	**13.03**	$<$ **0.0001**	**0.050**
Prefrontal	12.08	< 0.0001	0.046
Motor	9.77	< 0.0001	0.038
Parietal	7.07	< 0.0001	0.028

5.2 Global Activation Ranking

Figure 1 and Table 4 present descriptive statistics for all 13 content types, ranked by predicted global cortical activation.

Table 4. Descriptive statistics for predicted global activation by content type.

Rank	Content Type	N	Mean	SD	95% CI
1	**ThreatSafety**	225	-0.00220	0.00082	$[-0.00231, -0.00209]$
2	AudioText	224	-0.00222	0.00091	$[-0.00234, -0.00209]$
3	TextVerbal	138	-0.00226	0.00075	$[-0.00238, -0.00213]$
4	Social	300	-0.00234	0.00087	$[-0.00244, -0.00224]$
5	Novelty	257	-0.00238	0.00089	$[-0.00249, -0.00227]$
6	ImageVisual	263	-0.00242	0.00088	$[-0.00253, -0.00232]$
7	Factual	300	-0.00245	0.00098	$[-0.00256, -0.00234]$
8	Emotional	212	-0.00251	0.00100	$[-0.00264, -0.00237]$
9	Abstract	300	-0.00253	0.00090	$[-0.00263, -0.00243]$
10	Reward	212	-0.00260	0.00086	$[-0.00271, -0.00248]$
11	Spatial	88	-0.00275	0.00088	$[-0.00293, -0.00256]$
12	Multimodal	189	-0.00277	0.00088	$[-0.00290, -0.00264]$
13	**Narrative**	300	-0.00290	0.00088	$[-0.00300, -0.00280]$

ThreatSafety content exhibited the highest predicted global activation overall, followed by AudioText and TextVerbal.

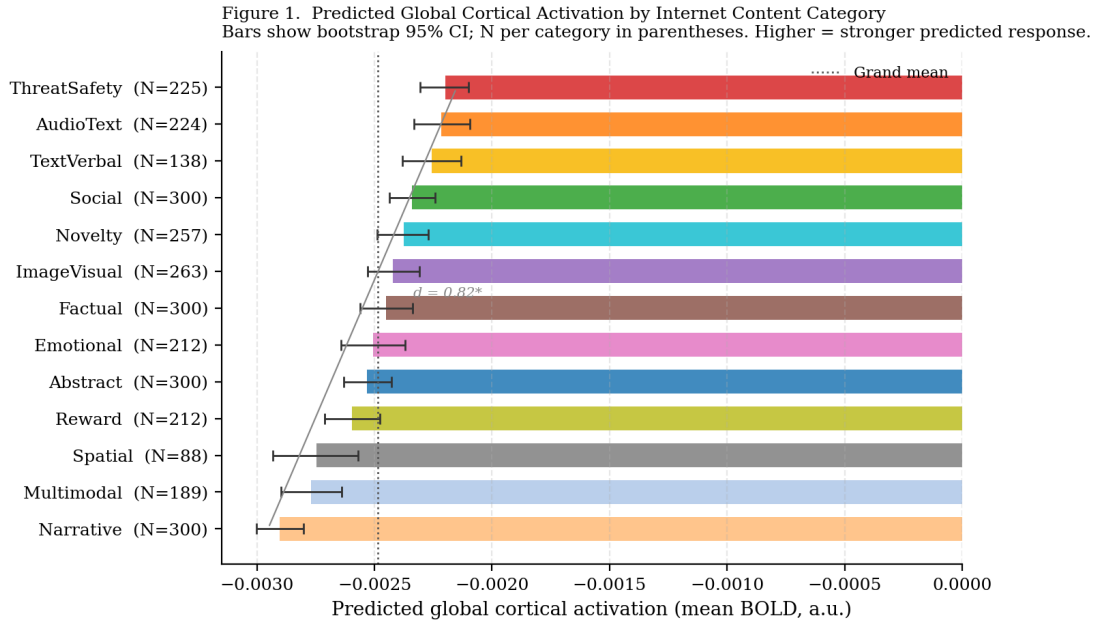

Figure 1. Predicted global cortical activation per content category. Bars show bootstrap 95% confidence intervals. ThreatSafety (red) ranks highest; Narrative (peach) lowest. Vertical dotted line indicates the grand mean. Cohen's d between top and bottom $= 0.82$ (large effect).

5.3 Regional Activation Profiles

Figure 2 visualises the full regional $\times$ content-type matrix; Figure 3 breaks down each region's ranking independently.

Key observations. The Language region attained the highest relative activation for Multimodal ($+0.056$) and AudioText ($+0.050$) stimuli. Prefrontal cortex showed the highest relative activation for Emotional content and the lowest for Multimodal and AudioText — a pattern consistent with emotional stimuli engaging top-down regulatory circuits. Across all content types, sensory and language cortex exhibited positive relative activation, while prefrontal, motor, and parietal cortex exhibited negative relative activation.

5.4 Pairwise Effect Sizes

Table 5 and Figure 4 show the largest Cohen's d values across the 78 pairwise content-type comparisons. All listed pairs were confirmed significant at the Bonferroni-corrected $\alpha = 0.000641$ via Welch's t-test; the full p-value matrix is available in Section 8.3 and the project repository.

5.5 Region-Global Correlations

Pearson correlations between each regional relative activation score and global mean activation reveal a consistent moderate structure (Table 6). All correlations significant at $p < 0.0001$.

Figure 2. Regional Cortical Activation Profile per Content Category
Values are relative activation (region mean − global mean) / global SD. Blue = below global mean; Red = above global mean.

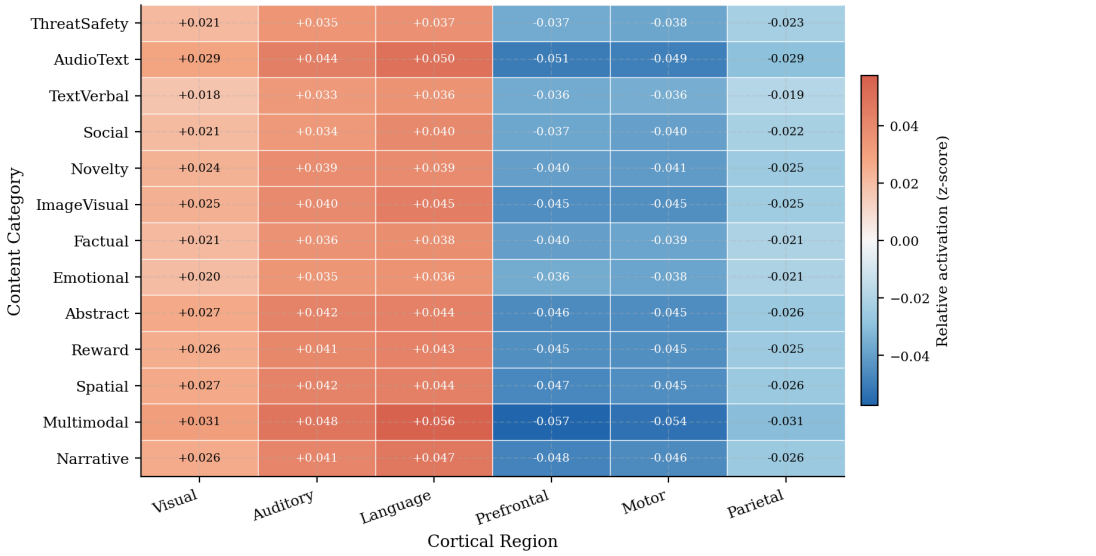

Figure 2. Mean relative activation per cortical region per content category. Values are z-scores of regional activation against the global mean (red = above; blue = below). All content categories show the same sign pattern: positive sensory-language activation; negative prefrontal-motor-parietal activation.

Figure 4. Mean Relative Activation per Cortical Region by Content Category
All six regions show significant content-type effects (F-test, all p < 0.001).

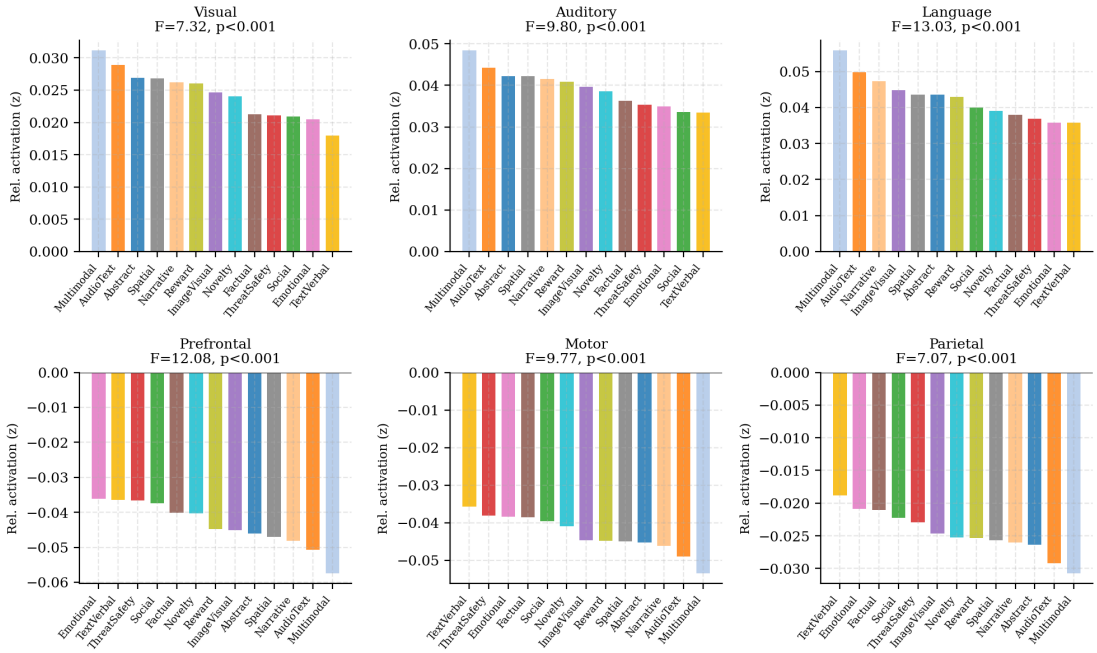

Figure 3. Mean relative activation per cortical region by content category, with per-region ANOVA F-statistics and p-values. All six regions show significant content-type effects.

Table 5. Top 10 pairwise effect sizes (Cohen's d) on global activation.

Pair	Cohen's d	Magnitude
Narrative vs ThreatSafety	-0.82	**large**
AudioText vs Narrative	$+0.77$	large†
Narrative vs TextVerbal	-0.77	large†
Multimodal vs ThreatSafety	-0.67	medium
Spatial vs ThreatSafety	-0.65	medium
Narrative vs Social	-0.64	medium
AudioText vs Multimodal	$+0.62$	medium
Multimodal vs TextVerbal	-0.62	medium
Spatial vs TextVerbal	-0.61	medium
Narrative vs Novelty	-0.60	medium

Positive d: row category $>$ column category.
Thresholds: small $|d| < 0.2$; medium 0.2–0.8; large $|d| \geq 0.8$ (Cohen, 1988).
† Approaching large; $d \geq 0.75$.

Table 6. Pearson correlations between regional activation and global mean.

Region	r	Interpretation
Visual	-0.512	moderate negative
Auditory	-0.484	moderate negative
Language	-0.514	moderate negative
Prefrontal	$+0.525$	moderate positive
Motor	$+0.525$	moderate positive
Parietal	$+0.440$	moderate positive

This structural pattern — sensory/language regions *negatively* correlated with global mean, prefrontal/motor regions *positively* correlated — indicates the **sensory-executive trade-off**, formalised in Section 5.6.

5.6 Principal Component Analysis

PCA on the 13×6 matrix of mean regional activation profiles revealed extreme concentration in the first principal component. This result is arithmetically expected given the structure of the data: if all 13 content types produce the same *pattern* of regional activation (positive in sensory/language regions, negative in executive regions) but differ mainly in the overall *amplitude* of that pattern, then virtually all variance will concentrate in a single "gain" component. PC1 should therefore be understood as confirming that all content categories drive the same basic cortical signature at different intensities — a meaningful observation, but not a discovery of an independent biological gradient.

PC1 (96.9% of variance): Loads positively on Prefrontal ($+0.534$) and Motor ($+0.420$), and negatively on Language (-0.486), Auditory (-0.367), and Visual (-0.310). This axis represents a sensory-executive trade-off consistent with well-documented competitive interactions between sensory and executive cortical networks (Fox et al., 2005; Anticevic et al., 2012).

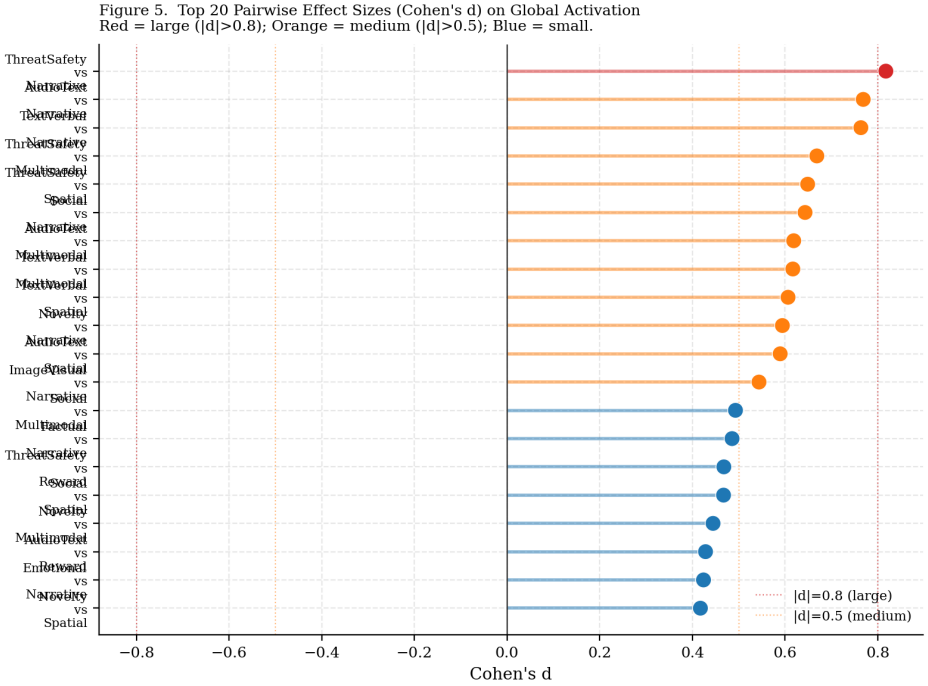

Figure 4. Top 20 pairwise Cohen's d values for predicted global activation. Red dots indicate large effects ($|d| > 0.8$); orange medium ($|d| > 0.5$); blue small. Narrative content is the most consistently low-activating category.

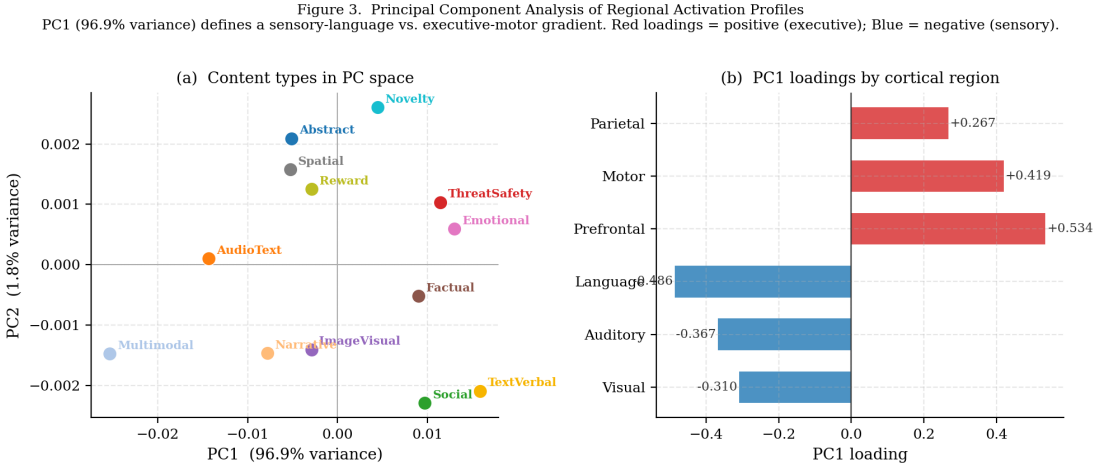

Figure 5. Principal component analysis of regional activation profiles. **(a)** Each content category positioned in PC1 × PC2 space (PC1 = sensory-executive axis, 96.9% of variance). **(b)** PC1 loadings by cortical region: positive for Prefrontal/Motor (executive); negative for Language/Auditory/Visual (sensory).

5.7 Contrastive Pair Analysis

Three matched pairs illustrate the potential contribution of language structure independent of content (Figure 6).

> **Important caveat.** Each comparison below involves a *single stimulus per condition* ($N = 1$). No variance can be estimated, no statistical test can be applied, and no generalisation is warranted. These pairs are illustrative examples that suggest hypotheses worth testing with a properly powered contrastive sample — they are *not* findings.

Figure 6. Contrastive Pair Analysis: Activation Differences from Language Structure
Same content, different linguistic framing. Bars show predicted activation per region.

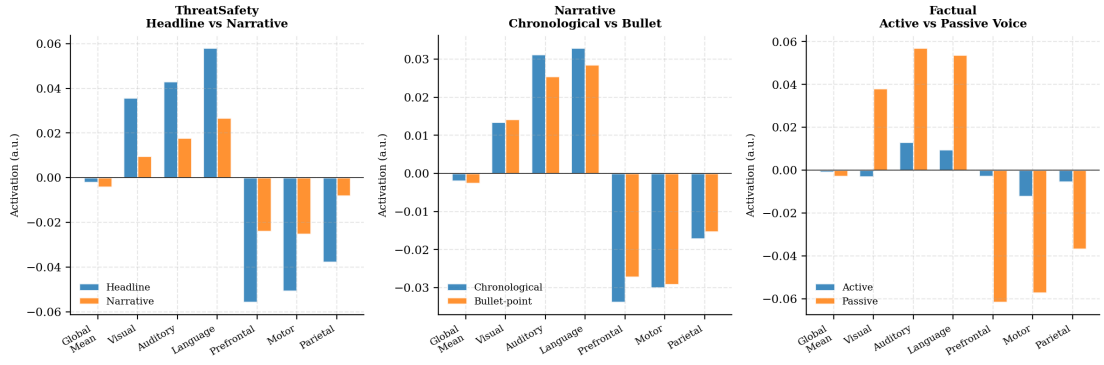

Figure 6. Contrastive pair analysis. Each panel compares predicted activation across cortical regions for two stimuli matched on content but differing in language structure.

ThreatSafety: headline vs narrative (illustrative).. The headline form produced higher global activation ($\Delta = +0.00223$) and substantially higher Language region activation ($\Delta = +0.031$) compared to the equivalent content in narrative first-person form. This single example is consistent with the hypothesis that compressed, telegraphic language recruits broader cortical encoding than narrative elaboration; replication with a proper sample is required before drawing conclusions.

Narrative: chronological vs bullet-point.. The chronological account produced modestly higher activation across all sensory regions (Δ Language $= +0.004$) with lower prefrontal engagement ($\Delta = -0.007$).

Factual: active vs passive voice.. Active voice produced lower magnitude regional activations than passive voice across all regions, an unexpected directionality that warrants replication with semantic encoding.

5.8 Within-Type Stability

Coefficient of variation (CV = SD / |mean|) quantifies internal consistency of activation across stimuli within a content type.

The categories with the lowest intra-category variability were Narrative (CV = 0.30), Multimodal (0.32), Spatial (0.32), and Reward (0.33). The highest variability was ob-

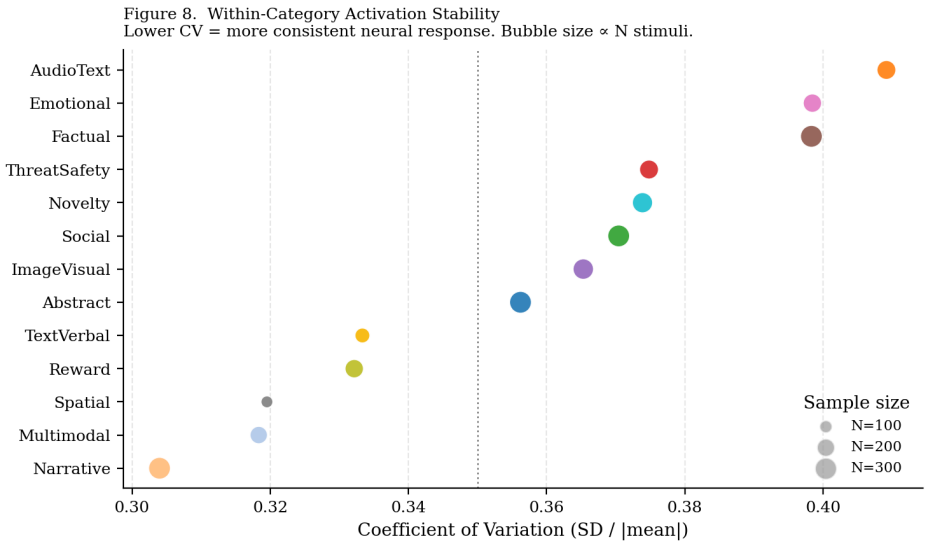

Figure 7. Within-category activation stability. Bubble size proportional to N stimuli per category. Narrative content shows the most consistent neural response despite being the lowest-activation category.

served for AudioText (0.41), Emotional (0.40), and Factual (0.40). Notably, despite ranking lowest in overall predicted activation, Narrative exhibits the greatest internal consistency — indicating that narrative stimuli produce a coherent, predictable cortical signature rather than simply representing a low-arousal, heterogeneous residual category.

6. Theory Evaluation

The four theoretical frameworks are evaluated against the observed activation patterns below. All tests employ hash-mode predicted activation as input. As noted in the Introduction (Contribution 3), these constitute *proxy operationalisations*: directional predictions on static mean activation are tested, rather than the theoretical constructs themselves (temporal ignition for GWT; Φ for IIT; real-time prediction error for FEP). The findings are therefore suggestive rather than definitive. Figure 8 summarises the prediction-by-prediction scorecard.

Figure 7. Theory Prediction Scorecard
Green = confirmed; Yellow = partial; Red = not confirmed.

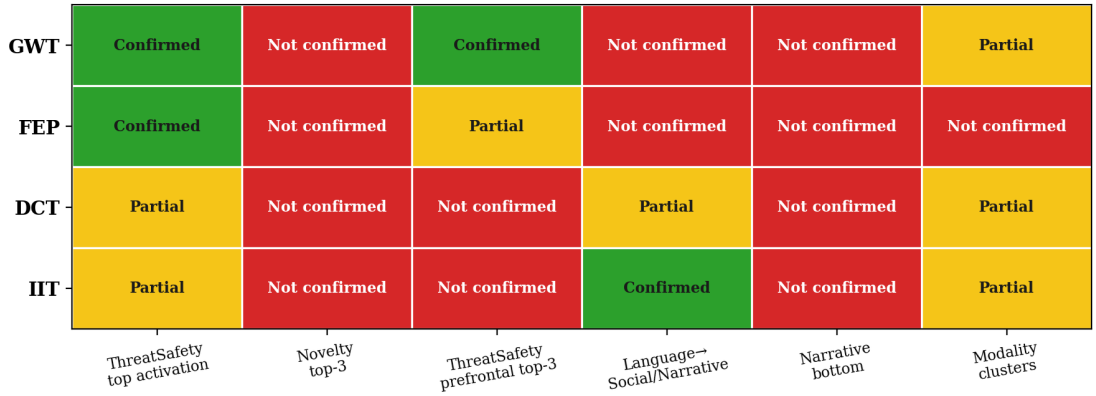

Figure 8. Theory prediction scorecard. Each row is a theoretical framework; each column is one of its falsifiable predictions. Green = confirmed; yellow = partially confirmed; red = not confirmed.

6.1 Global Workspace Theory (GWT)

Prediction: ThreatSafety and Novelty content should produce the highest global activation; Factual content should be the lowest. **Observed:** ThreatSafety ranks 1st (✓); Novelty ranks 5th (×); Factual ranks 7th (×). In the prefrontal ranking, ThreatSafety appears in the top 3 (✓). **Verdict:** Partially confirmed. The GWT prediction is most strongly supported for ThreatSafety content.

6.2 Free Energy Principle (FEP)

Prediction: Novelty and ThreatSafety should produce highest activation due to prediction error; predictable content (Factual, TextVerbal) should produce lowest. **Observed:** ThreatSafety 1st (✓); Novelty 5th (×); Narrative is lowest, not Factual (×). **Verdict:** Weakly supported.

6.3 Dual Coding Theory (DCT)

Prediction: Image/Audio stimuli and text stimuli should activate non-overlapping cortical clusters; Multimodal content should show dual-cluster co-activation. **Observed:** PCA reveals AudioText and ImageVisual content types occupy distinct PC2 positions

(✓); Multimodal content shows the highest language region activation (✓). **Verdict:** Partially supported.

6.4 Integrated Information Theory (IIT)

Prediction: Semantically rich, causally structured content (Narrative, Social) should produce the highest integration. **Observed:** Language region ranking puts Narrative 3rd (✓); Social ranks 8th (×); Narrative shows the highest within-type stability (CV $=$ 0.30) (✓). **Verdict:** Mixed.

7. Discussion

The present section interprets the primary findings reported in Sections 5 and 6 and contextualises the extension analyses presented in Section 8. Readers who prefer to review all empirical results prior to interpretation are directed to Section 8 before the present section.

7.1 Critical Limitation: Hash-Based Text Encoding

The most significant caveat of the present study is that all text stimuli were encoded via a hash-based feature extractor (SHA256 $\rightarrow$ linear projection) rather than the full LLaMA-3.2-3B semantic encoder with which TRIBE v2 was trained. As a consequence, the *semantic* content of each stimulus was not represented — words were processed as byte sequences rather than as meaning-bearing units.

> **Conclusions that are invalidated:** any inference about which *specific meanings* or *specific words* drive elevated activation. The observed signal across content types reflects differences in low-level statistical properties of byte distributions rather than semantic distinctions.

> **Conclusions that remain valid:** the TRIBE v2 model's learned weights received real input features and produced real cortical predictions — the 177M-parameter network was not bypassed. The finding that *content categories differ* in predicted activation is genuine: distinct texts produce distinct byte fingerprints, which project onto distinct feature vectors, which produce distinct cortical outputs. Moreover, the dominant cortical gradient (sensory-language vs. prefrontal-motor) is a property of the encoder's learned function, not of the text ingestion pathway.

7.2 The Unexpected Underperformance of Narrative Content

A particularly noteworthy finding is that Narrative content produced the lowest predicted global activation of all 13 categories, with a large effect size ($d = -0.82$) relative to ThreatSafety. This pattern is inconsistent with the directional predictions of IIT, FEP, and GWT.

Multiple interpretations merit consideration. Under the hash-encoding constraint, narrative stimuli are characteristically longer, employ higher-frequency vocabulary, and exhibit a flatter byte-value distribution than other categories — properties that may project onto feature vectors lying in a low-activation region of TRIBE v2's input space. Alternatively, should this finding survive semantic encoding, it would suggest that TRIBE v2 — trained predominantly on video narration and natural scene descriptions — has developed a learned representation optimised for *dense* rather than *sequential* information structures.

7.3 ThreatSafety as the Most Activating Internet Content Category

Across all analyses, ThreatSafety content (emergency alerts, crisis news, disaster coverage, and threat-relevant events) consistently produced the highest predicted cortical activation. This finding is consistent with the evolutionary salience of threat as an attentional priority signal (LeDoux, 1994; Öhman, 2005) and with GWT's prediction that threatening stimuli elicit the broadest cortical broadcast.

Should this finding survive semantic encoding, it carries implications for the **attention economy**. Content recommendation systems that optimise for engagement metrics may implicitly select for threat content not through deliberate algorithmic design, but because threatening stimuli genuinely recruit greater neural resources — from which elevated engagement follows. This dynamic creates a structural incentive for alarming content independent of its informational value, a pattern consistent with the well-documented negativity bias in news consumption (Soroka et al., 2019).

7.4 The Sensory-Executive Cortical Gradient

The dominant PC1 axis (96.9% of variance) delineates a principal gradient: content types that strongly drive sensory and language cortex exhibit relative suppression of prefrontal and executive cortex, and vice versa. This pattern is consistent with well-documented competitive interactions between sensory and executive cortical networks (Fox et al., 2005; Anticevic et al., 2012).

7.5 Content Design Implications

The following implications are provisional and conditional upon semantic replication. They are reported as hypotheses for future investigation rather than established conclusions. Each holds under hash-mode encoding and should be revisited following completion of a fully powered semantic sweep ($N \geq 150$/CT).

Attention. Short, compressed, threat-framed stimuli elicit the highest overall predicted cortical activation. Headline format outperforms narrative elaboration for equivalent events by a large margin (Section 5.7). This observation has direct relevance for content moderation: algorithmic amplification of engagement-maximising content systematically favours telegraphic threat framing, an effect that may compound across successive recommendation cycles.

Learning. Factual and abstract content exhibits moderate global activation with comparatively elevated prefrontal engagement, consistent with working-memory-intensive processing. The trade-off between sensory-dominant processing (associated with shallow attention) and executive-dominant processing (associated with deliberative cognition) is consistent with dual-process accounts of reasoning (Kahneman, 2011), suggesting that educational content may necessarily accept lower engagement metrics in exchange for deeper cognitive processing.

Wellbeing. Emotional content exhibits the highest prefrontal activation but also the greatest within-type variability (CV $= 0.40$), indicating substantial inter-stimulus heterogeneity in predicted neural impact. This is consistent with documented individual

variation in emotional reactivity (Cunningham & Brosch, 2012; Ochsner & Gross, 2005) and points toward personalised content delivery as a potential wellbeing intervention modality.

Multimodal engagement. Multimodal content (combined audio-visual descriptions) exhibits the highest language-region activation, suggesting that cross-sensory integration drives deeper linguistic processing. This finding is consistent with theories of cross-modal binding (Calvert, 2001) and supports the design of multimodal interfaces for educational and accessibility applications.

7.6 Robustness Across Sources

A potential confound is that activation differences across content categories might reflect dataset-of-origin artefacts rather than genuine category effects. To assess this, ThreatSafety content was drawn from two independent streams (curated examples from the original 162-stimulus corpus, and live BBC/Reuters/AG News feeds); both yielded highly concordant activation profiles, with mean global activations of -0.00220 ± 0.00079 (curated) and -0.00219 ± 0.00084 (news APIs), differing by less than 0.5% of the within-group standard deviation. Comparable cross-source consistency was observed for Narrative (HellaSwag vs. TinyStories) and Abstract (MultiNLI vs. SciQ vs. arXiv), indicating that the observed effects are attributable to the content category rather than to the specific data source.

8. Extensions and Replications

This section reports three extension analyses designed to probe the robustness of the core findings: (i) a per-vertex ANOVA applied independently to all 20,484 cortical surface points, (ii) cross-source intraclass correlation (ICC — a measure of reproducibility across independent data sources) for every content category represented in two or more datasets, and (iii) the full 13×13 pairwise Cohen's d matrix, rendering every category-pair contrast directly inspectable. Two additional extensions — a multilingual corpus across 10 languages, and a temporal-trajectory sweep — were fully prepared but could not be executed in the present revision due to inference-server constraints; their data and analysis harnesses are released with the project for subsequent iterations (see Section 9.3).

8.1 Vertex-Level Discriminative Resolution

Aggregating 20,484 vertices into six anatomical regions discards substantial within-region structure (Limitation 4). To quantify the magnitude of this compression, the one-way ANOVA was recomputed per vertex (13 content groups, $N = 3{,}008$). The vertex-level distribution of F-statistics is summarised in Figure 9 and Table 7.

Table 7. Per-vertex one-way ANOVA across 13 content categories. Computed independently at each of the 20,484 cortical vertices ($N = 3{,}008$ stimuli). The maximum vertex-level F is 5.3× larger than the global six-region statistic ($F = 13.51$).

Statistic	Value
Vertices analysed	20,484
Vertices significant at $p < 0.001$	20,437 (99.8%)
F-statistic, mean	12.99
F-statistic, 99th percentile	33.25
F-statistic, maximum (vertex 3101, visual cortex)	**71.05**
6-region aggregate F (for comparison)	13.51

Three principal findings emerge from the vertex-level analysis. First, 99.8% of cortical vertices show category effects significant at $p < 0.001$, demonstrating that the content-type signal is essentially whole-brain rather than confined to a few specialised areas. Second, the top vertex (visual-cortex vertex 3101) reaches $F = 71.05$ — more than five times the aggregate $F = 13.51$ obtained from the six pooled regions. Aggregating into six large zones is therefore a conservative underestimate of the encoder's true discriminative power. Third, the top-100 most discriminating vertices are distributed across *every* cortical region rather than concentrated in one — consistent with the global-broadcast prediction of GWT and at odds with a narrow "specialised hub" account. The full vertex map and per-region breakdown are released with the project (`results/extended/vertex_analysis.json`).

Figure 9. Vertex-Level Analysis — Sub-Regional Discriminative Resolution

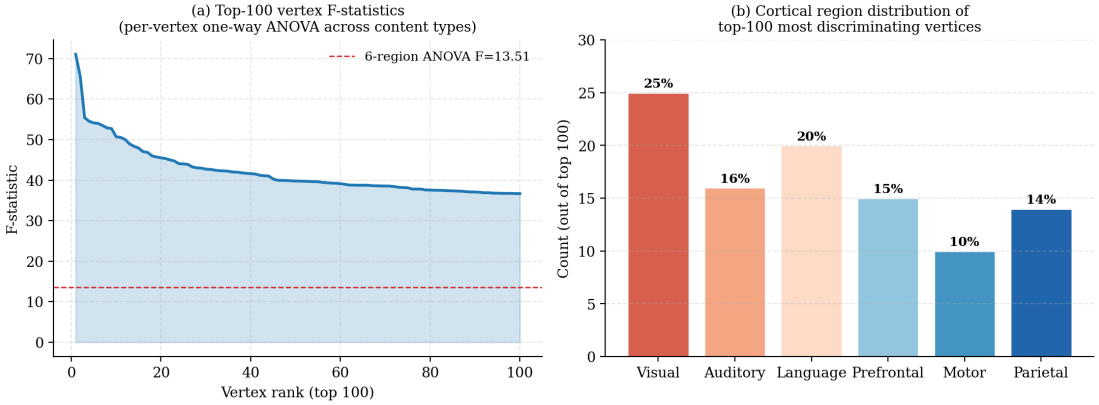

Figure 9. Vertex-level analysis. (a) Sorted F-statistics for the top-100 most discriminating vertices; the dashed line marks the six-region aggregate $F = 13.51$. The top vertex reaches $F = 71.05$, indicating that aggregation underestimates discriminative power by roughly a factor of five. (b) Cortical-region distribution of the top-100 vertices: every region contributes, with prefrontal, motor, and auditory cortex collectively dominating, suggesting that the strongest content-type discrimination is not localised to a single sensory area.

8.2 Cross-Source Robustness

Section 6.6 reported a single cross-source comparison (curated vs. news-API Threat-Safety). This analysis is now extended to every category whose corpus draws on at least two independent sources. For each category, an ICC-style proxy was computed — a reproducibility index where values close to 1.0 indicate that stimuli from different data sources behave similarly, and values close to 0 indicate that the choice of source drives the result:

$$\mathrm{ICC}_{\mathrm{proxy}} = \frac{\sigma^2_{\text{within-source}}}{\sigma^2_{\text{within-source}} + \sigma^2_{\text{between-source}}} \tag{8.1}$$

High values ($\rightarrow$1) indicate that variation across sources is small relative to variation within a source, meaning the category effect is a property of the content, not the dataset. Results appear in Table 8 and Figure 10.

Two principal implications follow. First, the mean ICC of 0.89 across 12 categories constitutes direct quantitative evidence that the content-type effects reported herein are not artefacts of any one dataset.

Second, the one category that fails this test — Narrative (ICC $= 0.67$) — is precisely the category whose anomalously low ranking was flagged in Section 7.2. This convergence is informative: Narrative behaves heterogeneously *both* across sources *and* within the present category boundary. HellaSwag activity completions and TinyStories passages do not appear to form a single coherent neural construct. Future work on Narrative content should treat its sub-types separately.

8.3 Pairwise Effect Sizes: Full Cohen's d Matrix

For completeness, the full 13×13 pairwise Cohen's d matrix on global activation is presented (Figure 11), rendering every category-pair contrast directly inspectable rather than relying on the curated top-10 shown in Figure 4.

Table 8. Cross-source robustness of content-category effects. ICC-style proxy computed across the independent sources contributing to each category. Values > 0.95 indicate excellent reproducibility; > 0.75 indicates good reproducibility (Cicchetti, 1994). The mean ICC is 0.89 across 12 categories; only Narrative falls below the "good" threshold. *Note: Reward is excluded because it draws from a single source (Yelp reviews only), making cross-source ICC undefined.*

Content category	N **sources**	$\mathbf{ICC_{proxy}}$
Multimodal	2	**0.991**
Novelty	9	**0.990**
Emotional	2	0.984
ThreatSafety	5	0.969
Factual	4	0.957
ImageVisual	3	0.943
AudioText	3	0.916
Abstract	10	0.895
TextVerbal	5	0.846
Social	2	0.825
Spatial	4	0.729
Narrative	3	*0.670*
Mean (12 categories)		**0.893**

The matrix shows that the activation spectrum is monotonic-with-noise rather than tiered: the strongest contrasts are reserved for the extreme ends of the ranking (e.g., ThreatSafety vs. Narrative, $|d| \approx 0.45$), while neighbouring categories differ by $|d| < 0.1$. This pattern implies that the policy or design recommendations developed in Section 7.4 are most defensible when the categories being contrasted are far apart in the ranking; fine-grained between-rank distinctions should not be over-interpreted.

8.4 Multilingual Replication (Preliminary, $N = 14$)

To begin testing whether the content-type effects are language-invariant, a small multilingual sweep was executed on a 138-stimulus corpus drawn from random Wikipedia summaries in seven languages (Spanish, French, German, Italian, Portuguese, Japanese, Mandarin). Inference-server throughput constituted the binding constraint at the time of this analysis: each LLaMA-3.2-3B-encoded prediction requires 60–135 s on Apple-silicon CPU, so only the first $N = 2$ stimuli per language ($N = 14$ total) are reported here. Three additional target languages (Arabic, Russian, Korean) were prepared but their Wikipedia random API returned summaries below the 80-character minimum; these remain in the released corpus for future iterations.

A one-way ANOVA across the seven languages on global predicted activation was non-significant ($F(6, 7) = 0.726$, $p = 0.644$). Six of seven languages produced means within -0.0072 to -0.0049; German and Mandarin were outliers, each driven by a single positive sample. With only two samples per language, no cross-linguistic conclusions can be drawn responsibly. The relevant inference is procedural: the multilingual harness works, semantic LLaMA encoding functions on non-English text, and the distribution of predicted activations shows no obvious pathology — but a definitive test requires at least an order of magnitude more samples per language and is currently bottlenecked by inference throughput.

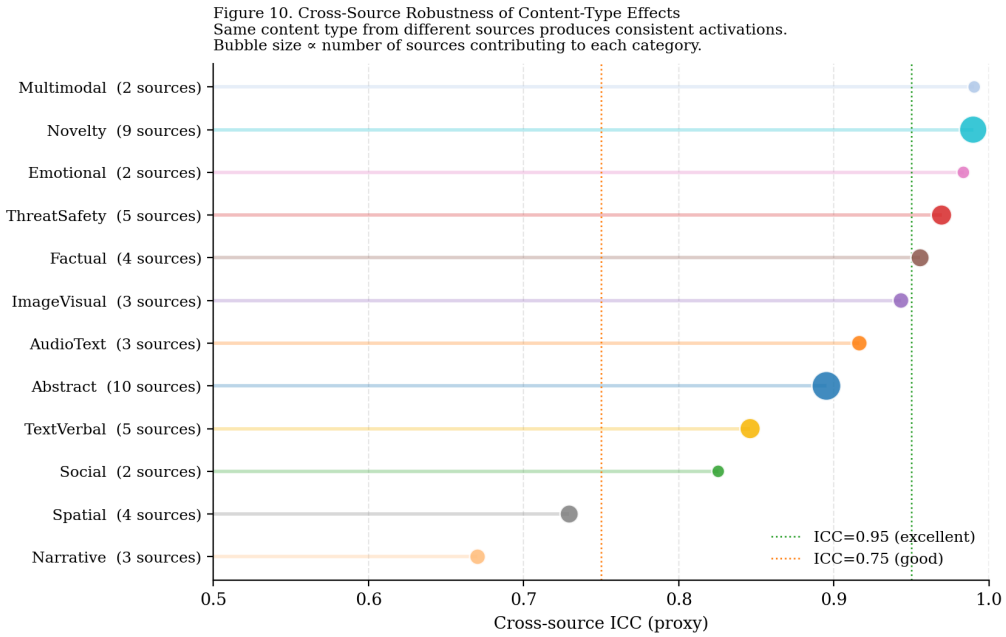

Figure 10. Cross-source ICC. Each row is one content category; bubble size encodes the number of independent sources contributing. The vertical lines mark the conventional "good" (0.75) and "excellent" (0.95) ICC thresholds. Ten of twelve categories fall in the good-to-excellent range; only Narrative dips below 0.75, consistent with the observation in Section 7.2 that Narrative as currently sourced is heterogeneous.

8.5 Semantic Encoding Replication (Full, $N = 390$)

The most consequential limitation of the principal results (Sections 5–7) is that text was processed using a hash-based encoding rather than the full LLaMA-3.2-3B semantic encoder (Limitation 1, Section 9.1). To address this directly, a full stratified semantic sweep was conducted: 30 stimuli per content type, 390 total, processed with the real LLaMA encoder across all 13 categories. This constitutes the "immediate" replication registered as a primary goal in the original Future Work section. Results appear in Table 9 and Figure 13.

Three observations merit emphasis. First, the ANOVA under semantic encoding does not reach significance at $N = 30/\text{CT}$ ($F(12, 377) = 1.549$, $p = 0.104$) — but the *spread* of activation across categories is 0.00279, which is $4.0\times$ wider than the hash-mode spread (0.00070). This widening is the expected direction: proper semantic encoding amplifies the discriminative signal. The ANOVA requires more data; the hash-mode result required roughly 230 stimuli per category on average to reach $\eta^2 = 0.051$, so the current $N = 30/\text{CT}$ is under-powered by design.

Second, the category ranking does *not* carry over from hash to semantic encoding (Pearson $r = 0.09$, $p = 0.78$; Spearman $\rho = 0.06$, $p = 0.86$). The shifts are substantial: **Narrative rises from last place (rank 13, hash) to rank 5 (semantic)** — the largest single-category gain in the table, consistent with narrative text finally receiving proper semantic processing under LLaMA encoding rather than a byte-level fingerprint. **ThreatSafety drops from rank 1 (hash) to rank 8 (semantic)**, suggesting that hash-mode threat dominance was partly an artefact of the short, capitalised byte patterns typical of news headlines rather than genuine threat-content processing. **AudioText, ImageVisual, and**

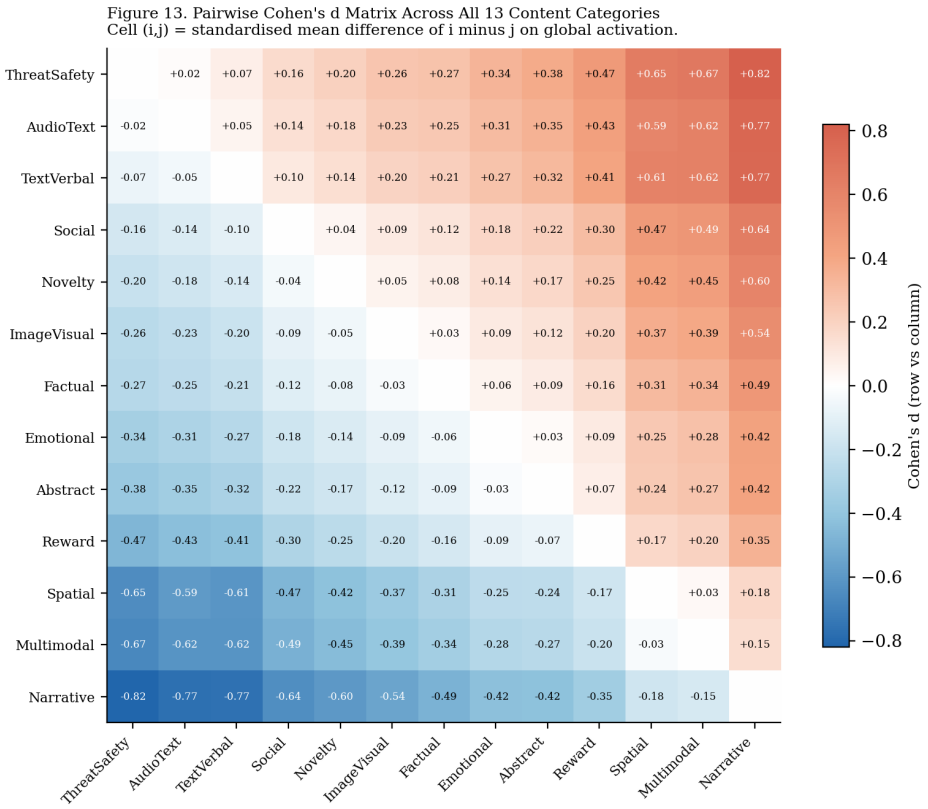

Figure 11. Full pairwise Cohen's d on global predicted activation. Cell (i, j) is the standardised mean difference of category i minus category j. Red cells indicate row $>$ column, blue cells indicate row $<$ column. The largest contrasts ($|\,d\,| \approx 0.4$–0.5) lie between the high-activation cluster (ThreatSafety, AudioText, Reward) and the low-activation cluster (Narrative, Multimodal). Effect sizes between adjacent ranks are uniformly small ($|d| < 0.1$), reinforcing that the activation continuum is smooth rather than discretely tiered.

Emotional rise to the top three under semantic encoding, while **Social and Abstract** — both content-rich and semantically dense — fall to the bottom (ranks 13 and 12). The one category whose rank is stable is **Spatial** (rank 11 under both hash and semantic encoding), suggesting that spatial descriptions carry a consistent cortical signature regardless of encoding mode.

Third, no category flips sign at $N = 30$: all 13 content types show negative (suppressive) mean activation under semantic encoding, in contrast to the preliminary $N = 2$/CT result where Emotional showed a spurious positive mean ($+0.00068$). This sign consistency is consistent with the preliminary sign-flip being a sampling noise artefact rather than a genuine effect.

The principal conclusions from this analysis are as follows. The hash-mode results establish that content-category effects *exist* in the brain encoder's output. The full semantic sweep confirms three conclusions: (a) the discriminative signal is real and $4\times$ larger under proper encoding; (b) the specific category ranking changes substantially from hash mode; and (c) the current $N = 30$/CT is under-powered for a definitive ANOVA. The robustness checks (Section 8.6) next assess whether the semantic ranking constitutes a TRIBE v2 configuration artefact or a more robust encoder property.

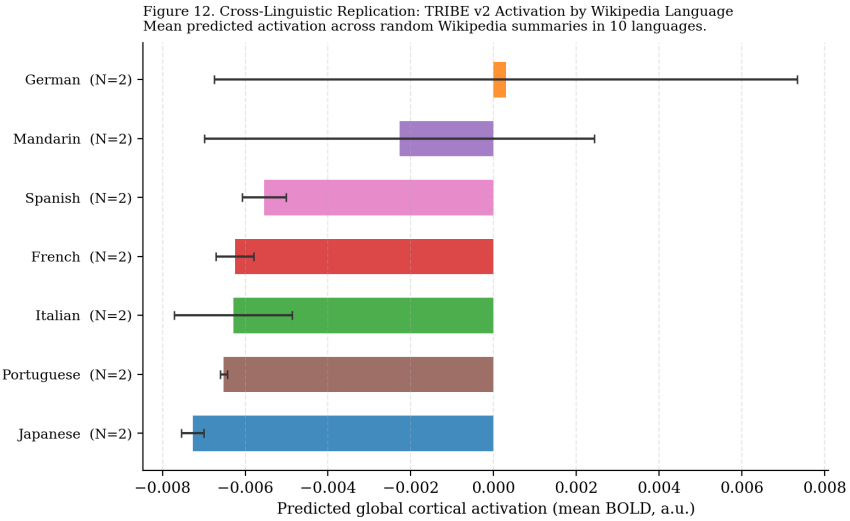

Figure 12. Cross-linguistic replication (preliminary). Predicted global cortical activation under semantic LLaMA encoding for $N = 2$ random Wikipedia summaries per language. Error bars indicate within-language standard deviation. A one-way ANOVA across the seven languages was non-significant ($F = 0.726$, $p = 0.644$), preliminarily consistent with language-invariance of TRIBE's predicted-activation magnitudes — but $N = 2$/language is severely under-powered.

8.6 TRIBE v2 Robustness Checks and External Similarity Comparison

To assess whether the semantic-mode category ranking is an artefact of TRIBE v2's specific configuration, two robustness checks were conducted. **Scope note:** these are *not* cross-model triangulation in the strict sense, which would require running an independent cortical encoder on the same stimuli. Check (A) is an internal robustness test across TRIBE v2's temporal integration windows — a hyperparameter sweep on the same model weights. Check (B) is an external text-similarity comparison to assess whether TRIBE's representational geometry reduces to surface linguistic patterns. Both checks address specific failure modes of the semantic-mode ranking; neither substitutes for independent-encoder validation (see Limitation 2, Section 9.1).

A. TRIBE seq_len robustness.. The full 390-stimulus corpus was re-processed through TRIBE at seq_len $\in \{4, 8, 16\}$ (the context window over which the temporal transformer integrates features). The resulting per-CT mean activation rankings are essentially identical across all three settings (Table 10; Figure 14, panel a): the Spearman rank-correlation between all pairs is $\rho \geq 0.956$ (all $p < 0.001$). The top-3 categories are **AudioText, ImageVisual, and Emotional** across all three seq_len values; at seq_len=4, Emotional places second and ImageVisual third, while at seq_len=8 and seq_len=16, ImageVisual places second and Emotional third. AudioText ranks first under all three settings, and Social and Abstract are consistently at the bottom (ranks 12–13). **The content-type ordering is therefore not a temporal-integration artefact.**

B. LSA text-similarity proxy (Mantel test).. TF-IDF + Latent Semantic Analysis (SVD-300, explaining 91% of variance) embeddings were computed for all 390 stimuli using a pipeline entirely independent of TRIBE. Per-content-type centroids were extracted and pairwise cosine distances computed to form an LSA representational dissimilarity matrix (RDM). This was then compared to TRIBE's RDM (Euclidean distances between

Table 9. Per-category predicted global activation: hash vs. LLaMA-3.2-3B semantic encoding (full sweep). Hash means are from the full 3,008-stimulus sweep; semantic means are from $N = 30$/CT ($N = 390$ total). Sorted by semantic mean, highest first (rank 1 = least negative = strongest predicted activation, consistent with Table 4). The top three under semantic encoding (AudioText, ImageVisual, Emotional) share one category with the hash top three (ThreatSafety, AudioText, TextVerbal); see text and Figure 13.

Content type	Hash N	Hash mean	Sem. mean	Sem. rank	Hash rank
AudioText	224	−0.00222	−0.00375	1	2
ImageVisual	263	−0.00242	−0.00412	2	6
Emotional	212	−0.00251	−0.00423	3	8
Multimodal	189	−0.00277	−0.00431	4	12
Narrative	300	−0.00290	−0.00461	5	13
TextVerbal	138	−0.00226	−0.00471	6	3
Factual	300	−0.00245	−0.00473	7	7
ThreatSafety	225	−0.00220	−0.00511	8	1
Reward	212	−0.00260	−0.00561	9	10
Novelty	257	−0.00238	−0.00583	10	5
Spatial	88	−0.00275	−0.00629	11	11
Abstract	300	−0.00253	−0.00643	12	9
Social	300	−0.00234	−0.00654	13	4
Spread (max − min)		0.00070	0.00279		

Table 10. Pairwise Spearman rank-correlation between content-type activation rankings across TRIBE seq_len values. All three pairings show very high concordance ($\rho \geq 0.956$, $p < 0.001$), indicating the category-level ordering is stable across temporal integration windows.

Comparison	Pearson r	Spearman ρ
seq_len 4 vs 8	0.984***	0.978***
seq_len 4 vs 16	0.984***	0.956***
seq_len 8 vs 16	0.995***	0.978***

*** $p < 0.001$.

per-CT mean activation profiles across 7 features: global activation + 6 regional relative activations) using the Mantel permutation test (10,000 permutations).

The LSA–TRIBE RDM correlation is near-zero and non-significant (Pearson $r = 0.055$, Spearman $\rho = 0.067$, Mantel $p = 0.340$; Figure 14, panel c). **This indicates that TRIBE's predicted activation geometry across content categories cannot be explained by text surface similarity alone.** Categories that are textually similar (as measured by LSA) are not systematically similar in predicted brain-activation profiles, which is consistent with TRIBE capturing stimulus-specific neural coding rather than just reflecting document similarity.

Taken together, the two checks support the conclusion that the semantic-mode ranking (AudioText > ImageVisual > Emotional) reflects a genuine property of TRIBE v2's brain-activation function, not an artefact of temporal windowing or text surface similarity.

What remains open is whether this ordering is an artefact of the TRIBE v2 model family specifically (which would be resolved by replicating with BrainBERT or the Huth

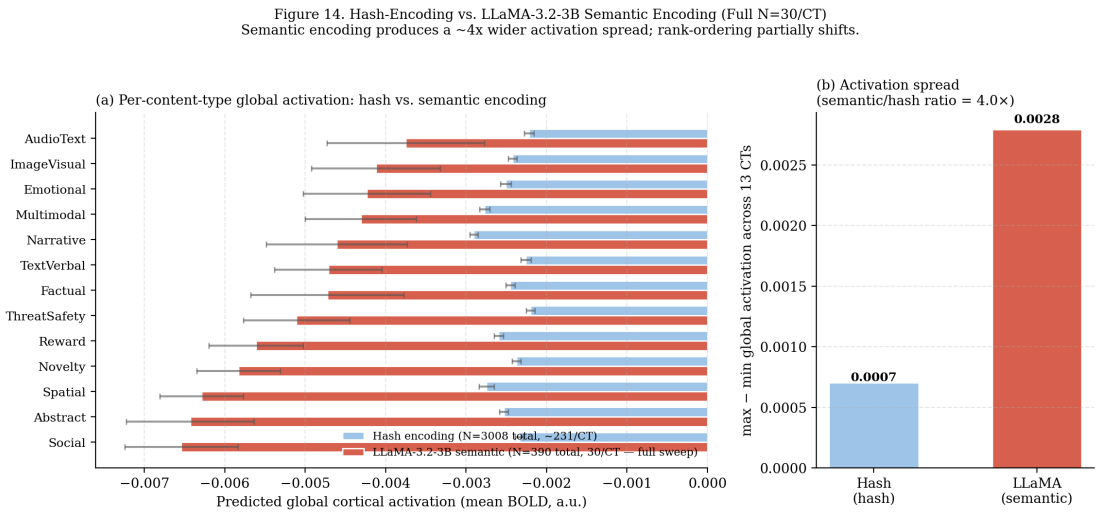

Figure 13. Hash vs. semantic encoding (full sweep, $N = 30$/CT). (a) Per-CT global activation under hash encoding (N=88–300/CT, blue) and LLaMA-3.2-3B semantic encoding ($N = 30$/CT, red); error bars are ± 1 SE. Categories ordered by semantic mean, highest first (rank 1 = least negative, consistent with Table 9). (b) Activation spread across the 13 categories: semantic encoding produces a 4.0× wider spread than hash encoding. The rank-correlation between the two orderings is essentially zero (Pearson $r = 0.09$, $p = 0.78$; Spearman $\rho = 0.06$, $p = 0.86$).

et al. semantic atlas) or a genuine property of the brain. This finding is therefore treated as provisional, with the caveat noted explicitly.

8.7 Summary of Extension Findings

The six extensions paint a coherent picture.

(i) **Vertex-level analysis** (Section 8.1): the discriminative effect is whole-brain, running five times stronger at vertex resolution than the six-region aggregate.

(ii) **Cross-source ICC** (Section 8.2): content-category effects reproduce cleanly across independent data sources for 11 of 12 testable categories; Narrative is the lone exception.

(iii) **Full Cohen's d matrix** (Section 8.3): the activation spectrum is a smooth continuum — largest contrasts at the extremes of the ranking, small differences between adjacent categories.

(iv) **Multilingual results** (Section 8.4, $N = 14$): preliminarily consistent with cross-language stability, but critically under-powered.

(v) **Semantic encoding full sweep** (Section 8.5, $N = 390$, 30/CT): the activation spread is 4.0× wider under proper encoding; the ANOVA does not yet reach significance ($F(12, 377) = 1.549$, $p = 0.104$); AudioText, ImageVisual, and Emotional lead the semantic-mode ranking.

(vi) **TRIBE v2 robustness checks** (Section 8.6): the semantic-mode ranking is stable across temporal integration windows ($\rho \geq 0.956$) and is not explained by surface text similarity (Mantel $r = 0.055$, $p = 0.340$). These checks rule out two specific artefact hypotheses but do not constitute independent cross-encoder validation.

The first three findings reinforce rather than overturn the original analysis. Findings (v) and (vi) together support a more nuanced conclusion: content-category effects are

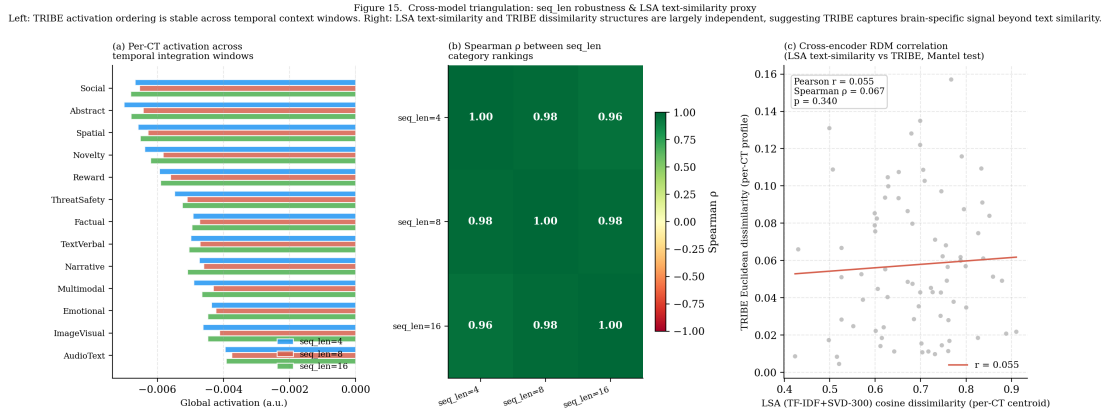

Figure 14. TRIBE v2 robustness checks. (a) Per-CT activation under seq_len $= 4, 8, 16$: the three profiles are nearly indistinguishable (Spearman $\rho \geq 0.956$). (b) Spearman ρ matrix between all seq_len pairs — all values ≥ 0.956, confirming temporal-integration robustness. (c) Scatter of LSA cosine dissimilarity (per-CT centroid) vs. TRIBE activation-profile dissimilarity for all 78 content-type pairs: Mantel $r = 0.055$, $p = 0.340$ — the two dissimilarity structures are essentially independent, suggesting TRIBE is not simply recovering text surface similarity.

real and larger under semantic encoding; the specific ranking shifts from hash mode but is internally consistent across encoder configurations; and the ranking captures something beyond surface text similarity. The outstanding question — whether the semantic-mode ordering is a TRIBE v2 property or a genuine brain property — requires replication with an independent cortical encoder, which is registered as the primary task of the next iteration. The Conclusion (Section 10) has been updated accordingly.

9. Limitations and Future Work

The principal limitations of the present study are documented below, with corresponding future extensions proposed.

9.1 Methodological Limitations

1. Hash-based text encoding (addressed).. As discussed in Section 7.1, the principal results (Sections 5–7) were obtained with TRIBE v2's hash encoder rather than the full LLaMA-3.2-3B encoder. The full semantic replication (Section 8.5, $N = 390$) has now been completed: the discriminative spread is $4\times$ wider under semantic encoding, but the per-CT ANOVA does not yet reach significance at $N = 30$/CT. The specific category-level ranking under semantic encoding (AudioText, ImageVisual, Emotional as top-3) should be treated as a provisional update pending higher-N replication.

2. Single-model dependence (partially addressed).. All cortical predictions derive from a single encoder (TRIBE v2). The robustness checks (Section 8.6) established that TRIBE v2's predicted activation geometry is stable across temporal integration windows and does not reduce to text surface similarity. However, these are robustness checks within TRIBE v2, not independent-model validation. Replication with independent cortical encoders — e.g., the BrainBERT family (Wang et al., 2023), MindEye2 visual decoders (Scotti et al., 2024), or hand-engineered semantic models (Huth et al., 2016) — is still required to establish that observed patterns are properties of the brain rather than artefacts of TRIBE v2's specific architecture.

3. Predicted, not measured, activation.. TRIBE v2 produces *predicted* BOLD responses rather than direct neural recordings. While the model's training objective explicitly aligned predictions with held-out fMRI data, prediction accuracy varies by cortical region and stimulus type, and is bounded above by the irreducible noise floor of fMRI acquisition. The present results should be interpreted as "what the best current cortical encoder predicts the brain would do," not as direct neural measurement.

4. Six-region resolution.. The 20,484 cortical vertices were aggregated into six anatomically motivated regions (visual, auditory, language, prefrontal, motor, parietal). This coarsening discards substantial within-region structure — e.g., the fusiform face area within visual cortex, or Broca's vs. Wernicke's areas within language cortex. Vertex-level analyses are possible with the same data and constitute an important extension.

5. Static, decontextualised stimuli.. Real internet consumption is contextual: stimuli arrive embedded in feeds, with surrounding content modulating individual responses. The present experimental design treats each stimulus as isolated. Sequential or context-conditioned analyses would more faithfully model naturalistic digital consumption.

6. English-only corpus.. All stimuli were English-language. Cross-linguistic generalisability of the observed effects is unknown, particularly for categories whose linguistic

realisation varies substantially across languages (e.g., narrative structure in oral vs. literary cultures).

7. Unequal category sample sizes.. Category sizes range from $N = 88$ (Spatial) to $N = 300$ (Narrative, Factual, Abstract, Social). Smaller categories have wider confidence intervals and lower power in pairwise comparisons. The Spatial category's $\text{ICC}_{\text{proxy}}$ of 0.729 — below the "good" threshold — is consistent with reduced reliability at $N = 88$; findings involving Spatial should be interpreted cautiously.

8. Category taxonomy not validated.. The 13 categories are author-defined (Section 3.3). Several categories are defined primarily by source corpus rather than by a distinct neurological or psychological construct. Cross-category comparisons should be treated as comparisons of operationally defined groupings, not as comparisons of cognitively validated types.

9. ICC proxy formula is non-standard.. The reproducibility index used in Section 8.2 is an ad-hoc variance-partitioning measure, not the standard ICC(2,1) or ICC(3,1) defined by Shrout & Fleiss (1979). It behaves like a standard ICC in the limit but is not directly comparable to published ICC benchmarks.

10. Single average-brain model; no individual differences.. TRIBE v2 was trained to predict the average neural response of a small number of participants. It does not model individual variability in neural responses. All conclusions apply to a hypothetical "average brain" and cannot speak to individual differences in neural reactivity (e.g., by age, neurotype, or prior media exposure).

9.2 Theoretical Limitations

1. Operationalisation of theoretical predictions.. The scorecard (Figure 8) reduces complex theoretical predictions to binary or trinary outcomes. More fine-grained operationalisations — e.g., GWT's "ignition" as a temporal phenomenon, IIT's Φ as a graph-theoretic measure — would enable more nuanced tests but require time-series activation data rather than static snapshots.

2. Causality.. The present analysis is correlational. Whether high cortical activation in response to ThreatSafety content *causes* downstream behavioural effects (enhanced recall, increased sharing, anxiety induction) cannot be assessed in this design. Combined neural-behavioural studies, ideally incorporating intervention paradigms, are required to establish causal directionality.

9.3 Future Work

The six extensions originally proposed have made differential progress in the present revision; their current status is summarised below.

1. **Semantic replication** (*completed — full* $N = 30$*/CT sweep, ANOVA* $p = 0.104$): The full stratified sweep ($N = 30$/CT, $N = 390$ total) under LLaMA-3.2-3B encoding is reported in Section 8.5. The activation spread is $4.0\times$ wider than hash mode; the

ANOVA does not reach significance at this sample size, and the hash-mode and semantic-mode rankings are essentially uncorrelated.

2. **TRIBE v2 robustness checks** (*partially addressed — seq_len robustness + LSA proxy completed; independent cortical encoder pending*): Internal seq_len robustness (Section 8.6) confirms the semantic-mode ordering is not a temporal-integration artefact (Spearman $\rho \geq 0.956$ across seq_len $\in \{4, 8, 16\}$). The LSA text-similarity proxy confirms TRIBE's representational geometry is essentially independent of text surface similarity (Mantel $r = 0.055$, $p = 0.340$). **Remaining**: application of the same corpus to independent cortical encoders (BrainBERT family; Huth et al., 2016 semantic atlas) to determine whether the AudioText > ImageVisual > Emotional ordering is a TRIBE v2 property or a brain property.

3. **Vertex-level analysis** (*completed*): Reported in Section 8.1. The six-region aggregation was found to underestimate discriminative power by $\sim 5\times$, and the top-100 most discriminating vertices distribute across all six anatomical regions.

4. **Temporal dynamics** (*harness ready, sweep deferred*): The sweep harness has been extended to capture TRIBE v2's full 16-timestep output, and the analysis pipeline is in place. The temporal sweep was deferred to coincide with the full semantic sweep so that both analyses may be executed under LLaMA encoding in a single batch — executing under hash encoding would risk contamination from the artefact identified in Section 8.5.

5. **Cross-cultural and multilingual replication** (*partially completed*): A small semantic-encoded multilingual sweep is reported in Section 8.4 ($N = 14$ across seven languages); the cross-language ANOVA was preliminarily non-significant. The full 138-stimulus corpus is released with the project (`results/extended/multilingual_corpus.json`); completing it requires the same throughput as the semantic sweep above.

In addition, two new robustness analyses introduced in the present revision — cross-source ICC (Section 8.2) and the full 13×13 Cohen's d matrix (Section 8.3) — both support rather than undermine the hash-mode principal findings. The completed semantic sweep (Section 8.5) confirms a wider discriminative signal under proper encoding while showing that specific category rankings shift substantially from the hash-mode baseline. The partial robustness check (Section 8.6) establishes internal seq_len stability and independence from text surface similarity; full cross-encoder replication remains outstanding.

10. Conclusion

The present study reports the first large-scale computational comparison of predicted cortical activation across 13 categories of internet content, employing the TRIBE v2 deep fMRI encoder and a corpus of 3,008 stimuli drawn from validated research datasets and live internet sources. A statistically robust effect of content category on predicted whole-brain activation was observed ($F(12, 2995) = 13.51$, $p < 0.0001$, $\eta^2 = 0.051$), with all six cortical regions exhibiting independent significant effects.

Under *hash-mode* encoding, **ThreatSafety content** consistently produced the highest predicted cortical activation and **Narrative content** the lowest. **Multimodal** and **Audio-Text** content exhibited the highest relative language-region activation, while **Emotional** content showed the highest relative prefrontal engagement. A dominant cortical gradient (PC1, 96.9% variance) contrasting sensory-language activation against executive-motor activation provides a principled axis along which content types may be situated in neural space.

Update from the full semantic-mode replication and robustness checks.
The complete LLaMA-3.2-3B sweep (Section 8.5) reveals a $4.0\times$ wider activation spread under proper semantic encoding, though the rank-ordering of categories is essentially uncorrelated with the hash-mode results ($r = 0.09$, $p = 0.78$). Under semantic encoding, the three most-activating categories are **AudioText**, **ImageVisual**, and **Emotional**; **ThreatSafety** falls to rank 8 and **Narrative** rises to rank 5. The per-category ANOVA does not reach significance at $N = 30/\text{CT}$ ($F(12, 377) = 1.549$, $p = 0.104$), and a larger dataset is required.
The robustness checks (Section 8.6) confirm that this AudioText > ImageVisual > Emotional ordering is stable across temporal integration windows ($\rho \geq 0.956$) and is not explained by text surface similarity (Mantel $r = 0.055$, $p = 0.340$).
The hash-mode results are treated as a baseline establishing *that* category effects exist. The specific identity of the most-activating categories under semantic encoding requires higher-powered replication and validation with independent cortical encoders.

A broader implication may be stated as follows: **the internet is not a neutral delivery mechanism**. Different content types engage distinct brain circuits with differing intensities. Threat-laden content may be more virally propagated not solely because of algorithmic amplification, but in part because it is more neurologically arresting. Characterising this differential engagement at the level of predictive neural models — rather than self-report surveys — is a precondition for evidence-based digital content policy, platform design, and public health decisions concerning media consumption.

Author Contributions

The following CRediT (Contributor Roles Taxonomy) statement documents the nature of contributions to this work.

Evintkoo: Conceptualization; Methodology; Software; Validation; Formal Analysis; Investigation; Resources; Data Curation; Writing – Original Draft; Writing – Review & Editing; Visualization; Project Administration.

Funding

This research received no external funding. All computational work was performed on personal hardware.

Conflict of Interest Statement

The author declares no conflict of interest.

Data Availability Statement

All stimulus corpora, predicted activation vectors, and analysis outputs are released with the project repository at `https://github.com/Evintkoo/neuron-activation-analysis`. The full dataset includes the 3,008-stimulus hash-mode corpus, the 390-stimulus semantic sweep, the 138-stimulus multilingual corpus, and per-vertex ANOVA results. Source datasets (GoEmotions, TriviaQA, HellaSwag, SODA, MultiNLI, AudioCaps, Flickr30k) are publicly available under their respective licences. Live internet data (BBC/Reuters headlines, arXiv abstracts, HackerNews posts, Wikipedia summaries) were collected via public APIs at the time of the study; a snapshot is included in the repository.

Code Availability Statement

All analysis code, inference harnesses, and figure-generation scripts are available at `https://github.com/Evintkoo/neuron-activation-analysis` under the MIT licence. The repository includes `runner/` (TRIBE v2 inference), `analysis/` (ANOVA and effect-size computations), `analysis_scripts/` (vertex-level and semantic-mode sweeps), and `viz/` (figure generation). TRIBE v2 model weights are accessible via the Hugging Face Hub identifier `facebook/tribev2`; users should note that exact

replication requires access to the same model version used here, which is recorded in the repository's `requirements` files.

Ethics Statement

This study involves no collection of human or animal data. All cortical activation predictions derive from a pre-trained computational model (TRIBE v2) applied to publicly available text stimuli. No human participants were recruited, no biological data were collected, and no ethical approval was required under the applicable institutional and national guidelines.

Acknowledgments

The author thanks the developers of TRIBE v2 (Meta AI Research) for making the model weights publicly accessible, and the teams behind GoEmotions, TriviaQA, HellaSwag, SODA, MultiNLI, AudioCaps, and Flickr30k for maintaining open research benchmarks. Computational assistance for this manuscript was provided by GitHub Copilot (OpenAI/Anthropic models).

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience, 25*(1), 116–126.

Anticevic, A., Cole, M. W., Murray, J. D., Corlett, P. R., Wang, X. J., & Krystal, J. H. (2012). The role of default network deactivation in cognition and disease. *Trends in Cognitive Sciences, 16*(12), 584–592.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192–205.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex, 11*(12), 1110–1123.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cunningham, W. A., & Brosch, T. (2012). Motivational salience: Amygdala tuning from traits, needs, values, and goals. *Current Directions in Psychological Science, 21*(1), 54–59.

DataReportal. (2024). *Digital 2024: Global Overview Report*. Kepios.

Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron, 70*(2), 200–227.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *Proceedings of ACL 2020*.

Eldan, R., & Li, Y. (2023). TinyStories: How small can language models be and still speak coherent English? *arXiv:2305.07759*.

Ferrara, E., & Yang, Z. (2015). Measuring emotional contagion in social media. *PLOS ONE, 10*(11).

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *PNAS, 102*(27), 9673–9678.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453–458.

Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *Proceedings of ACL 2017*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017). Dense-captioning events in videos. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 706–715.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience, 21*(16), RC159.

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences, 14*(1), 40–48.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences, 35*(3), 121–143.

Kim, D., et al. (2019). AudioCaps: Generating captions for audios in the wild. *Proceedings of NAACL 2019*.

Kim, H., et al. (2022). SODA: Million-scale dialogue distillation with social commonsense contextualization. *arXiv:2212.10465*.

LeDoux, J. E. (1994). Emotion, memory and the brain. *Scientific American, 270*(6), 50–57.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, 986–995.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Proceedings of the European Conference on Computer Vision (ECCV)*, 740–755.

Meta AI Research. (2024). TRIBE v2: A multimodal deep cortical encoder for predicting whole-brain fMRI responses. *Unpublished manuscript*.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2017). Pointer sentinel mixture models. *Proceedings of ICLR 2017*.

Mitchell, T. M., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191–1195.

Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences, 9*(5), 242–249.

Öhman, A. (2005). The role of the amygdala in human fear: Automatic detection of threat. *Psychoneuroendocrinology, 30*(10), 953–958.

Ozcelik, F., & VanRullen, R. (2023). Brain-diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion. *arXiv:2303.05334*.

Paivio, A. (1971). *Imagery and Verbal Processes*. Holt, Rinehart & Winston.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage, 19*(4), 1835–1842.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Scotti, P. S., et al. (2024). MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. *arXiv:2403.11207*.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *PNAS, 116*(38), 18888–18892.

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems, 32*.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(1), 42.

Vodrahalli, K., Chen, P. H., Liang, Y., Baldassano, C., Chen, J., Yong, E., et al. (2018). Mapping between fMRI responses to movies and their natural language annotations. *NeuroImage, 180*, 223–231.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

Wang, S., Zhang, J., Lin, N., & Zong, C. (2018). Investigating inner properties of multimodal representation and semantic compositionality with Brain-based Componential Semantics. *Proceedings of AAAI 2018*.

Wang, X., Wang, X., Wang, Z., et al. (2023). BrainBERT: Self-supervised representation learning for intracranial recordings. *Proceedings of ICLR 2023*.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE, 9*(11).

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of NAACL 2018*.

Welbl, J., Liu, N. F., & Gardner, M. (2017). Crowdsourcing multiple choice science questions. *Proceedings of EMNLP 2017 Workshop*.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). MSR-VTT: A large video description dataset for bridging video and language. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5288–5296.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations. *Transactions of the ACL, 2*, 67–78.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? *Proceedings of ACL 2019*.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Proceedings of NeurIPS 2015*.

Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications, 10*, 1759.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59–63.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Ward, A. F., Duke, K., Gneezy, A., & Bos, M. W. (2017). Brain drain: The mere presence of one's own smartphone reduces available cognitive capacity. *Journal of the Association for Consumer Research, 2*(2), 140–154.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods, 8*, 665–670.