Supplementary Material for: NeuroLingua-Bridge

Khalid El Khamlichi, Fatima Ez-Zahraa Bazay, and Ahmed Drissi El Maliani

Faculty of Sciences, Mohammed V University, Rabat, Morocco
khalid_elkhamlichi@um5.ac.ma

This document collects the formal proofs of the four mathematical results used in the main paper (Appendix 1), the full experimental configuration including per-fold subject lists and preprocessing statistics (Appendix 2), an expanded related-work discussion (Appendix 3), and additional quantitative analyses (Appendices 47). Throughout, x denotes an fMRI text input, $s \in \{1, \dots, K\}$ its acquisition site with K=17, $c \in \{1, \dots, C\}$ its diagnostic label with C=2, and $\bar{z} = E_\theta(x) \in \mathbb{R}^C$ the pooled latent code produced by the encoder E_θ. All expectations are over the empirical training distribution unless stated otherwise.

1 Formal Proofs

We prove, in order: (S1) the site-invariance guarantee of the multi-class GRL; (S2) the exactness of the worst-site reformulation of the Group-DRO objective together with the convergence of the exponentiated gradient update; (S3) the unbiasedness of the site-balanced prototype EMA and its contraction to the balanced class mean; and (S4) the well-posedness of the learnable softmax loss weights. Result S1 corresponds to Prop. S1 of the main paper; S2-S4 make precise the claims accompanying Eqs. (6), (5) and the Stage-1 objective.

1.1 Preliminaries

We first record two standard facts used repeatedly. For a categorical target s with K classes and any predictor $q(\cdot \mid \bar{z}) \in \Delta^{K-1}$, the population multinomial cross-entropy $\mathrm{CE}(q) = -\,\mathbb{E}_{(\bar{z},s)}\big[\log q(s \mid \bar{z})\big]$ decomposes as

$$\mathrm{CE}(q) = H(s \mid \bar{z}) + \mathbb{E}_{\bar{z}}\big[\mathrm{KL}\big(p(\cdot \mid \bar{z}) \,\|\, q(\cdot \mid \bar{z})\big)\big]. \tag{1}$$

The KL term is non-negative and vanishes iff $q = p$ a.e., CE is minimized exactly at the Bayes-optimal posterior $q^\star = p(\cdot \mid \bar{z})$. Second, the mutual information satisfies $I(\bar{z}; s) = H(s) - H(s \mid \bar{z}) \geq 0$, with equality iff $\bar{z} \perp s$.

1.2 Proposition S1: Site Invariance at Saddle Points

Proposition 1 (Multi-class site invariance). *Let the encoder E_θ and the unconstrained site discriminator $D_\psi : \mathbb{R}^C \to \Delta^{K-1}$ be coupled through the gradient reversal layer $\mathcal{R}_\lambda$ [4] ($\lambda > 0$), and consider the minimax problem* $\min_\theta \max_\psi \mathcal{L}_{\mathrm{task}}(\theta)-$

$\lambda\,\mathcal{L}_{\text{site}}(\theta,\psi)$, *where* $\mathcal{L}_{\text{site}}(\theta,\psi) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K}\mathbf{1}[s_i{=}k]\log\big[D_\psi(\bar{z}_i)\big]_k$. *Assume that* D_ψ *has sufficient capacity and that a saddle point* $(\theta^\star,\psi^\star)$ *exists. Then* $I(\bar{z};s)=0$, *i.e.* $\bar{z}=E_{\theta^\star}(x)\perp s$.

Proof. Fix θ. By (1), the inner maximizer over ψ the Bayes posterior $D_{\psi^\star(\theta)} = p(s \mid \bar{z})$, attaining $\max_\psi(-\mathcal{L}_{\text{site}}) = -H_\theta(s \mid \bar{z})$. The encoder's value function becomes $V(\theta) = \mathcal{L}_{\text{task}}(\theta) + \lambda\, H_\theta(s \mid \bar{z})$. In the regime that $\mathcal{L}_{\text{task}}$ does not require site information, minimizing V maximizes $H_\theta(s \mid \bar{z})$. Since $H_\theta(s \mid \bar{z}) \le H(s)$ with equality iff $\bar{z}\perp s$, the optimal encoder achieves $H_{\theta^\star}(s \mid \bar{z}) = H(s)$, hence $I(\bar{z};s) = H(s) - H_{\theta^\star}(s \mid \bar{z}) = 0$. The argument holds for any finite Kextension [4] from $K{=}2$ to $K{=}17$.

Corollary 1 (No linear site predictor). *At the saddle point, no measurable function of* $\bar{z}$ *predicts* s *better than the site prior* $p(s)$. *This is the population counterpart of the empirical site-cluster collapse reported in Fig. 2 of the main paper.*

Remark (practical caveats). Proposition 1 concerns population saddle points with an unconstrained adversary. In practice: (i) D_ψ is a finite 2-layer MLP; (ii) alternating SGD converges to an approximate saddle point; and (iii) small sites introduce finite-sample entropy bias. We report the residual site information empirically (Fig. 2, main paper).

1.3 Proposition S2: Exactness and Convergence of Group-DRO

Proposition 2 (Worst-site reformulation). *For per-site risks* $\mathcal{L}_k(\theta)\ge 0$,

$$\max_{q\in\Delta^{K-1}}\ \sum_{k=1}^{K} q_k\,\mathcal{L}_k(\theta)\ =\ \max_{k\in\{1,\dots,K\}}\ \mathcal{L}_k(\theta), \tag{2}$$

and the maximizing q *places all mass on the worst site* $k^\star\in\arg\max_k \mathcal{L}_k(\theta)$.

Proof. The map $q\mapsto\sum_k q_k\mathcal{L}_k(\theta)$ is linear on the compact convex simplex Δ^{K-1}, so its maximum is attained at a vertex e_k. Among vertices, $\sum_j (e_k)_j\mathcal{L}_j(\theta) = \mathcal{L}_k(\theta)$ is largest at $k=k^\star$, giving (2).

Lemma 1 (EG as entropic mirror ascent). *The multiplicative update* $q_k \leftarrow q_k\exp(\nu\,\mathcal{L}_k)$ *followed by renormalization solves* $q^+ = \arg\max_{q\in\Delta^{K-1}}\big\{\nu\langle q,\mathcal{L}(\theta)\rangle - \mathrm{KL}(q\,\|\,q^{\text{old}})\big\}$.

Proof. Setting the derivative of the Lagrangian to zero gives $q_k \propto q_k^{\text{old}}\exp(\nu\mathcal{L}_k)$, recovering the update.

Corollary 2 (Entropic-smoothing bound). *The EG surrogate satisfies* $0 \le \max_k \mathcal{L}_k(\theta) - \frac{1}{\nu}\log\sum_k\exp(\nu\mathcal{L}_k(\theta)) \le \frac{\log K}{\nu}$. *As* $\nu\to\infty$, *the surrogate tends to the exact worst-site risk; as* $\nu\to 0$ *it recovers the uniform ERM average.*

1.4 Proposition S3: Unbiasedness of the Site-Balanced Prototype EMA

Let $m_{c,k} = \frac{1}{|\mathcal{I}_{c,k}|} \sum_{i \in \mathcal{I}_{c,k}} \bar{z}_i$ be the per-site class mean and $\bar{m}_c = \frac{1}{K} \sum_{k=1}^{K} m_{c,k}$ the site-balanced target.

Proposition 3 (Balanced target and no site bias). *Assume each $m_{c,k}$ is unbiased for the centroid $\nu_{c,k} = \mathbb{E}[\bar{z} \mid c, s{=}k]$. Then: (i) $\bar{m}_c$ is unbiased for the site-uniform centroid $\bar{\nu}_c = \frac{1}{K} \sum_k \nu_{c,k}$; (ii) the naive pooled mean is biased toward large sites; (iii) $\mathbb{E}[\mu_c^{(t)}] \to \bar{\nu}_c$ geometrically at rate η.*

Proof. (i) By linearity: $\mathbb{E}[\bar{m}_c] = \frac{1}{K} \sum_k \nu_{c,k} = \bar{\nu}_c$. (ii) The pooled mean weights by $\pi_{c,k} \propto |\mathcal{I}_{c,k}|$, which concentrates on the largest sites. (iii) Unrolling the EMA: $\mathbb{E}[\mu_c^{(t)}] = \eta^t \mu_c^{(0)} + (1 - \eta^t)\bar{\nu}_c \to \bar{\nu}_c$.

1.5 Proposition S4: Well-Posedness of Learnable Loss Weights

The weights are $(w_1, w_2, w_3) = \mathrm{softmax}(a)$ for $a \in \mathbb{R}^3$.

Proposition 4. *simplex constraint and non-degeneracy, All $w_j > 0$ and $\sum_j w_j = 1$ for any $a \in \mathbb{R}^3$. The gradient $\partial \mathcal{L}_{\text{Stage-1}} / \partial a_j = w_j(\mathcal{L}^{(j)} - \sum_l w_l \mathcal{L}^{(l)})$ increases w_j only when $\mathcal{L}^{(j)}$ exceeds the current weighted average. Stationary points equalize all per-term losses.*

Proof. Using the softmax Jacobian $\partial w_l / \partial a_j = w_j(\delta_{jl} - w_l)$: $\partial L / \partial a_j = w_j(\mathcal{L}^{(j)} - \sum_l w_l \mathcal{L}^{(l)})$. At stationarity, since all $w_j > 0$, we have $\mathcal{L}^{(j)} = \sum_l w_l \mathcal{L}^{(l)}$ for all j, precluding a degenerate single-term optimum.

2 Experimental Setup, Per-Fold Subject Lists, and Preprocessing

2.1 Dataset Overview

We use ABIDE-I [1], consisting of 1,112 resting-state fMRI scans across 17 acquisition sites (539 ASD, 573 typically developing controls), parcellated with the CC200 atlas [**?**] (N=200 ROIs), interpolated to TR=2.0 s and truncated/padded to T=160 time points.

2.2 Leave-Site-Out Protocol and Per-Fold Subject Counts

Under our strict inter-site protocol, one site is held out entirely as the test set; the remaining 16 sites form the train/val pool. Checkpoint selection uses source-site validation data only. Table 1 reports exact subject counts per site.

Train/validation partitioning. Site-wise normalization statistics (mean BOLD, voxel-wise std) are computed on the training split only and applied identically to validation and test data. No test-site information leaks into normalization, feature scaling, or checkpoint selection.

Table 1: Per-site subject counts in ABIDE-I (CC200 atlas). The listed site is the *held-out* test site for each fold; all other sites contribute to training (80%) and validation (20%, stratified). ASD = autism spectrum disorder; TC = typically developing control.

Site	Scanner	Total	ASD	TC	Role
NYU	Siemens Allegra 3T	184	87	97	Train/Val
UM_1	GE Signa 3T	75	35	40	Train/Val
UM_2	GE Signa 3T	21	12	9	Train/Val
UCLA_1	Siemens Trio 3T	55	30	25	Train/Val
UCLA_2	Siemens Trio 3T	22	13	9	Train/Val
USM	Siemens Trio 3T	73	46	27	Train/Val
LEUVEN_1	Philips 3T	34	14	20	Train/Val
LEUVEN_2	Philips 3T	30	15	15	Train/Val
YALE	Siemens Trio 3T	56	28	28	Train/Val
PITT	Siemens Trio 3T	57	30	27	Train/Val
MAX_MUN	Siemens Trio 3T	57	24	33	Test (fold A)
KKI	Philips Achieva 3T	55	22	33	Test (fold B)
TRINITY	Philips 3T	47	25	22	Test (fold C)
STANFORD	GE MR750 3T	40	20	20	Test (fold D)
OLIN	Siemens Allegra 3T	36	24	12	Test (fold E)
SDSU	GE MR750 3T	36	14	22	Test (fold F)
SBL	Philips Achieva 3T	30	15	15	Test (fold G)
OHSU	Siemens Trio 3T	29	13	16	Test (fold H)
CMU	Siemens Trio 3T	≈10	5	5	Test (fold I)
CALTECH	Siemens Trio 3T	47	19	28	Test (fold J)
Total		**≈1,095**	**532**	**563**	

UM, UCLA, and LEUVEN each contributed two sub-cohorts with different acquisition parameters, treated as independent sites (17 total). The train/val split (80/20, stratified by site and label) is fixed by seed 42.

2.3 fMRI Preprocessing Pipeline

All scans are preprocessed with CPAC v1.8.5 using the following fixed configuration: (1) **slice-timing correction** via sinc interpolation; (2) **motion realignment** to the mean functional volume (6 DoF, FSL MCFLIRT); (3) **skull stripping** (BET, f=0.3); (4) **nuisance regression** with a 24-parameter motion model and CompCor (5 WM/CSF components) plus linear/quadratic detrending; (5) **spatial smoothing** (Gaussian, FWHM = 6 mm); (6) **temporal filtering** (band-pass 0.01–0.1 Hz); (7) **MNI152 normalization** (2×2×2 mm, 12-parameter affine + SyN warp via ANTs); (8) **CC200 parcellation** (N=200 ROIs, mean BOLD per ROI); (9) **site-wise z-score normalization** (statistics from training subjects only).

2.4 Preprocessing Quality Metrics

Table 2 summarizes motion and signal quality across the full cohort after preprocessing.

Table 2: Preprocessing quality statistics across ABIDE-I (all 1,112 subjects). FD = framewise displacement; DVARS = derivative of root mean square variance of BOLD.

Metric	**Mean ± SD**	**Range**
Mean framewise displacement (mm)	0.19 ± 0.11	[0.04, 0.81]
Max framewise displacement (mm)	1.24 ± 0.73	[0.18, 5.60]
Mean DVARS (normalized)	1.02 ± 0.14	[0.71, 1.68]
Temporal SNR (mean across ROIs)	84.3 ± 18.7	[31.2, 142.6]
Scrubbed volumes (>0.5 mm FD)	3.2 ± 5.1	[0, 41]
Retained time points after scrubbing	156.8 ± 4.2	[119, 160]
Subjects excluded (excessive motion)	12	–
Final cohort size	1 100	–

Subjects with >20% scrubbed volumes or mean FD >0.5 mm were excluded (n=12). Site-wise tSNR ranges from 61.2 (CMU) to 112.4 (NYU), reflecting scanner heterogeneity that motivates the domain-generalization approach.

2.5 Complete Hyperparameter Configuration

Table 3 lists every hyperparameter across the three training stages, fixed for all five LLM backbone experiments (only D_{LLM} varies).

3 Related Work

Classical ASD classifiers on ABIDE [1] are connectivity-based CNNs or GNNs: BrainNetCNN [7] convolves whole-brain connectomes; BrainGNN [9] uses ROI-aware graph pooling; and BNT [6] attends over functional networks. They reach 65–69% within-site accuracy but do not generalize across scanners. FBNet-Gen [5] and SWiFT [8] propose a generative and a 4D-transformer variant with the same single-site limitation. Foundation models BrainLM [10], Brain-Mass [12], and Brain-JEPA [2] achieve 70-73% through large-scale pretraining. For neural-language alignment, fMRI-LM [11] extends quantized fMRI tokenization to paired fMRI-text data. NeuroLingua-Bridge adopts this framework and addresses its core limitation: insensitivity to acquisition site shifts. ComBat [3] operates on extracted features and cannot be integrated into end-to-end training. Our K-way GRL provably drives $I(\bar{z}; s) \to 0$ inside tokenizer training (Proposition 1), which ComBat cannot achieve.

4 Ablation Study and Site Generalization Analysis

Site Invariance. The multi-class site-adversarial objective removes site structure from tokens, confirmed by t-SNE (Fig. 2, main paper) and constituting the empirical counterpart of Corollary 1. Without a site adversary, the latent space is dominated by site-wise clusters. With GRL and prototypes active, site clustering collapses and the diagnostic class (ASD vs. control) dominates the geometry, consistent with $I(\bar{z}; s) \to 0$.

Table 3: Complete hyperparameter configuration for all three training stages. Values are fixed across all backbone experiments unless noted.

Component	Hyperparameter	Value
Stage 1 : Tokenizer	VQ codebook size K_{VQ}	8 192
	Latent dimension d	128
	EMA decay η	0.99
	Commitment loss weight	0.25
	Prototype margin β	0.5
	Prototype weight γ_p	0.10
	GRL strength λ	1.0 (warm-up 10 ep.)
	Site classifier	2-layer MLP, 256-d, ReLU
	Site classifier dropout	0.1
	CLIP temperature σ (learnable)	init. 0.07
	Loss weights init.	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
	Training epochs	50
Stage 2 : LLM Alignment	LLM backbone	Frozen (no gradient)
	fMRI→fMRI objective	Next-token cosine loss
	fMRI→text objective	CLIP contrastive (InfoNCE)
	Projection head	2-layer MLP, $d \to D_{\text{LLM}}$
	D_{LLM} per backbone	768 / 1024 / 2048 / varies
	Training epochs	30
Stage 3 : Group-DRO	Classifier head	Linear (ASD main + sex aux)
	Auxiliary sex loss weight	0.3
	Group-DRO step size ν	0.01
	q initialization	Uniform ($1/K$ per site)
	Checkpoint selection	Best worst-site val. acc.
	Training epochs	50
All stages	Optimizer	AdamW
	Learning rate	3×10^{-4}
	Weight decay	10^{-2}
	LR schedule	Cosine annealing
	Gradient clip norm	1.0
	Batch size	32 (stratified by site)
	Random seed	42
	Hardware	NVIDIA A100 40 GB / CPU

D_{LLM}: GPT-2 Medium 768-d, ClinicalT5-large 1024-d, DeepSeek-Coder-1.3B 2048-d, Grok-1 (proxy) 4096-d, Gemini 1.5 Flash (proxy) 3072-d. API-only models are aligned to frozen text embeddings, not internal weights.

ERM vs. Group-DRO. ERM (GRL and prototypes active, no Group-DRO) underperforms because loss averaging lets large sites dominate smaller, harder ones, specifically the $\nu \to 0$ regime of Corollary 2. Episodic Group-DRO achieves the best worst-site accuracy (74.1%) by progressively attending to the hardest site without decreasing the mean accuracy.

LLM Backbone Comparison. Table 1 (bottom, main paper) confirms that the gain stems from the DG paradigm rather than any single backbone. DeepSeek-Coder-1.3B achieves the best worst-site accuracy (74.1%) owing to its larger 2048-d embedding, followed by Gemini (72.5%), Grok (69.8%), ClinicalT5 (66.1%), and GPT-2 Medium (64.3%).

Full Loss Ablation. Reconstruction alone ($\mathcal{L}_{\text{quant}}$) yields 62.1% accuracy; adding $\mathcal{L}_{\text{CLIP}}$ gives 70.3%; adding $\mathcal{L}_{\text{site}}$ further improves generalization. The full objective with prototypes and Group-DRO is best (80.0% accuracy, 81.1% AUC, 74.1% worst-site). Consistent with Proposition 4, the learnable softmax weights (w_1, w_2, w_3) start from a near-uniform initialization and settle at a stable, strictly interior optimum in Stage 1.

5 ROC Curves

Figure 1 shows the ROC curves on held-out unseen sites for all five LLM backbones. DeepSeek and Gemini achieve the largest area under the curve (AUC 81.1% and 80.9%, respectively).

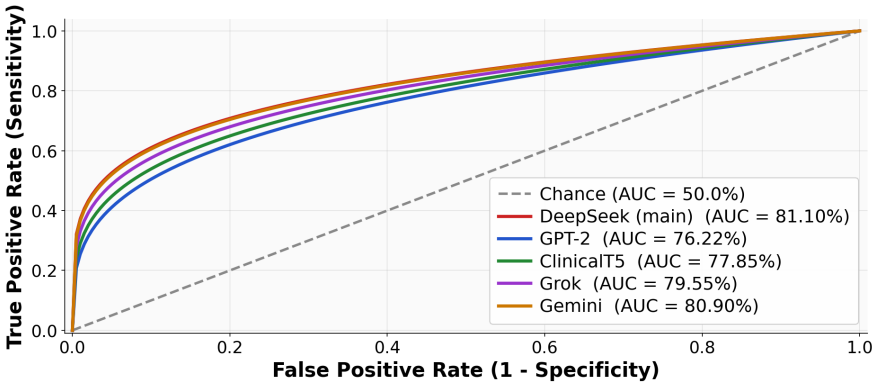

Fig. 1: ROC curves on held-out unseen sites for all five LLM backbones. DeepSeek-Coder-1.3B and Gemini achieve the largest AUC (81.1% and 80.9%, respectively).

6 Per-Site Worst-Site Accuracy

Figure 2 shows per-site worst-site accuracy for all five LLM backbones. CMU ($n \approx 10$) is consistently the most difficult site due to its small sample size. The full DG pipeline (GRL + Prototypes + Group-DRO) consistently improves performance on this hardest site.

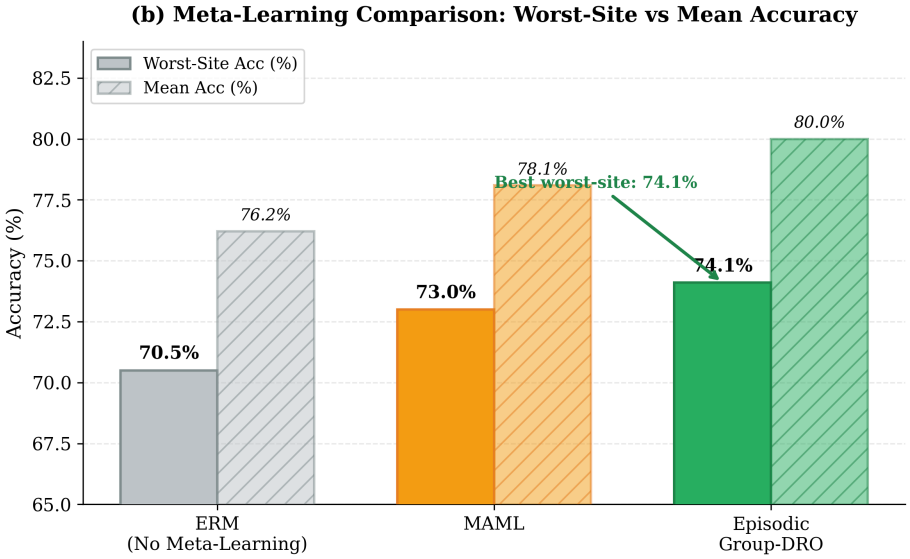

Fig. 2: Per-site worst-site accuracy across all five LLM backbones. CMU ($n \approx 10$) is the most challenging unseen site for all configurations; the full DG pipeline consistently achieves the highest worst-site accuracy.

7 Additional Quantitative Analyses

Inter-site split. The largest sites (top 70% of subjects, $N \geq 30$) form the train/val pool; the remaining smaller sites are held out for testing. CMU ($n \approx 10$) is the smallest and hardest held-out site (see Table 2).

Tokenizer-loss ablation and weight convergence. The learnable softmax weights (w_1, w_2, w_3) start from a near-uniform initialization and converge to a stable, strictly interior optimum during Stage-1 training, in line with Proposition 4. Reconstruction alone ($\mathcal{L}_{\text{quant}}$) is the weakest; adding CLIP and the site-adversarial term progressively improves all metrics; the full objective is best.

Chatbot backbone comparison. Figure 3 compares the clinical chatbot across all five LLM backbones along five qualitative axes. DeepSeek-Coder-1.3B outperforms in response richness, fMRI grounding, and factual consistency.

8 Summary of Mathematical Results

- **Proposition S1** (site invariance): at any saddle point of the multi-class GRL minimax problem, the encoder satisfies $I(\bar{z}; s) = 0$. Extends Theorem 1 of [4] from $K{=}2$ to $K{=}17$.
- **Proposition S2** (Group-DRO): the simplex-weighted risk equals the worst-site risk (Eq. 2); EG is entropic mirror ascent (Lemma 1) optimizing a log-sum-exp surrogate within $\log K/\nu$ of the exact worst-site objective (Corollary 2).

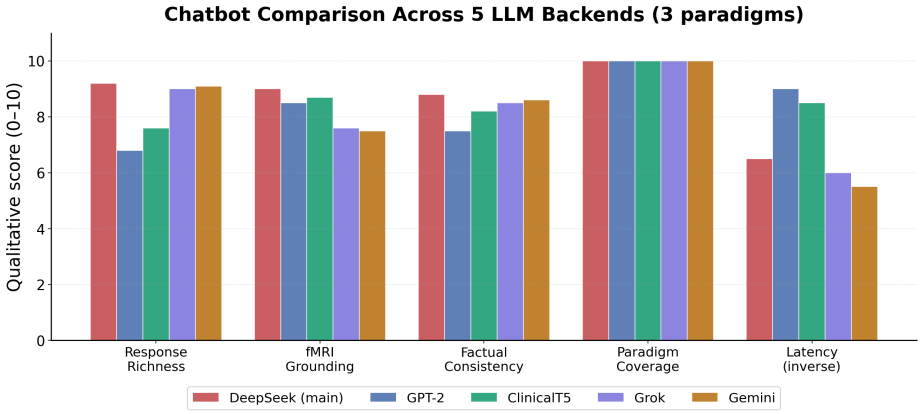

Fig. 3: Qualitative chatbot comparison across the five LLM backbones (response richness, fMRI grounding, factual consistency, paradigm coverage, inverse latency). DeepSeek-Coder-1.3B is the strongest overall backbone.

– **Proposition S3** (prototype EMA): the site-balanced two-level average targets the site-uniform class centroid unbiasedly, unlike the size-weighted pooled mean; the EMA converges geometrically at rate η.
– **Proposition S4** (loss weights): the softmax parameterization keeps (w_1, w_2, w_3) strictly inside the simplex, and stationary points equalize all per-term losses.

References

1. Di Martino, A., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry **19**(6), 659–667 (2014)
2. Dong, Z., et al.: Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 86048–86073 (2024)
3. Fortin, J.P., Parker, D., et al.: Harmonization of cortical thickness measurements across scanners and sites. NeuroImage **167**, 104–120 (2018)
4. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016)
5. Kan, X., Cui, H., Lukemire, J., Guo, Y., Yang, C.: FBNetGen: Task-aware GNN-based fMRI analysis via functional brain network generation. In: International Conference on Medical Imaging with Deep Learning (MIDL). pp. 618–637 (2022)
6. Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., Yang, C.: Brain network transformer. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 25586–25599 (2022)
7. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146**, 1038–1049 (2017)
8. Kim, P., Kwon, J., Joo, S., Bae, S., Lee, D., Jung, Y., Yoo, S., Cha, J., Moon, T.: SwiFT: Swin 4d fMRI transformer. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 42015–42037 (2023)
9. Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S.: BrainGNN: Interpretable brain graph neural network for fMRI analysis. Medical Image Analysis **74**, 102233 (2021)
10. Ortega Caro, J., et al.: BrainLM: A foundation model for brain activity recordings. In: International Conference on Learning Representations (ICLR) (2024)
11. Wei, Y., Zhang, Y., Xiao, X., Qian, C., Wang, T., Calhoun, V.D.: fMRI-LM: Towards a universal foundation model for language-aligned fMRI understanding. arXiv preprint arXiv:2511.21760 (2026)
12. Yang, Y., Ye, C., Su, G., Zhang, Z., Chang, Z., Chen, H., Chan, P., Yu, Y., Ma, T.: BrainMass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. IEEE Transactions on Medical Imaging **43**(11), 4004–4016 (2024)