CNeuroMod Data Collection Complete: 200h of individual fMRI Across Diverse Naturalistic and Controlled Tasks to build NeuroAI models

Julie A. Boyle (julieaboyle@criugm.qc.ca)
Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, Canada

Basile Pinsard (basile.pinsard@gmail.com)
Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, Canada

Lune (Pierre) Bellec (lune.bellec@umontreal.ca)
Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, Canada
Department of Psychology, University of Montreal, Montréal, Canada
Mila, University of Montreal, Montréal, Canada

& The CNeuroMod team
https://docs.cneuromod.ca/en/latest/AUTHORS.html

Abstract

Dense fMRI datasets are gaining popularity as a powerful resource to build integrative NeuroAI models to understand the brain. The Courtois Project on Neuronal Modelling (CNeuroMod) has now completed a 5-years data collection process resulting in 200 hours of fMRI data per subject using diverse naturalistic and controlled cognitive tasks. CNeuroMod is the largest dense fMRI dataset currently available to support the development of individualized and generalizable AI models of complex brain processes.

Keywords: dense fMRI dataset, multimodal datasets, NeuroAI, individual brain models, Algonauts 2025

Introduction

Several large individual fMRI datasets have emerged to train artificial intelligence (AI) models on specific cognitive processes like natural image (NSD: Allen et al. 2021; BOLD5000: Chang, et al. 2019) and movies (Dr Who: Seeliger et al. 2019) viewing. However, a key feature of the brain is the capacity to integrate and switch between specialized processes and cognitive contexts.The CNeuroMod project, which just completed its data collection, is the largest individual fMRI dataset to date. We collected a rich neuroimaging dataset that probes numerous cognitive domains in the same subjects (N=6) using both carefully controlled and engaging naturalistic tasks in order to build versatile and complex Neuro-AI models. Additionally, the overlap between CNeuroMod tasks and other open data resources opens the possibility to test NeuroAI models transferability and generalization across subjects and datasets.

Methods

Participants (aged 31 to 47 at the time of recruitment in 2018); 3 women and 3 men; 3 native French & 1 native English speakers, 2 bilingual; all right-handed) consented to participate for at least 5 years of data collection. All participants had good general health and normal hearing for their age. Four subjects were scanned approximately 80h / year and two were scanned approximately 40h / year. FMRI data were acquired with a 3T scanner). Setup included physiological signal recording, headcases to minimize motion, and a custom-built fiber optic controller to play videogames (Harel et al., 2023). Data were preprocessed with the fMRIprep pipeline LTS (Esteban et al, 2019).

The CNeuroMod Databank

The databank includes 29 fMRI datasets targeting several cognitive domains and modalities. Per subject, there are 197h of fMRI data, 17h of anatomical MRI data (Boudreau et al, 2025), 29h of data collected outside the scanner, including longitudinal hearing measures (Fortier et al, 2025) and videogame (Shinobi) training. The following is a breakdown of the fMRI datasets by cognitive domains, per subject (Ss), Table 1).

Vision. 22 datasets (175h) including movies (10h), 7 seasons of TV-show Friends (70h), & functional localizers (Stigliani et al, 2015; Kay et al. 2013). **Language.** 19 datasets (73h), including listening to Le Petit Prince (3h; Li et al, 2022) in 3 languages, reading a chapter of harrypotter (Wehbe et al 2014), triplets is a semantic association task involving triplets of words (7h), functional localizers (Scott et al, 2017 & Malik-Moralda et al 2021). **Memory.** 5 datasets (44h), including a memory task (18h) using THINGS (18h; Hebart et al. 2019), and multfs (8h), which is a working-memory task. **Emotions.** 13 datasets (147h), including gifs (5h) that evoke varying emotional dimensions (Cowen & Keltner, 2017). **Auditory.** 18 datasets (146h), including Mutemusic, which is an auditory imagery task, and narratives, a listening task (Nastase et al., 2018) with verbal recal inside the scanner. **Videogames (**48h). In-scan playing of videogames including Shinobi, SuperMario, Mariostars, and Mario3.

OOD Algonauts (2h/Ss)**.** A secret dataset acquired for the Algonauts 2025 challenge (Gifford et al., 2025).

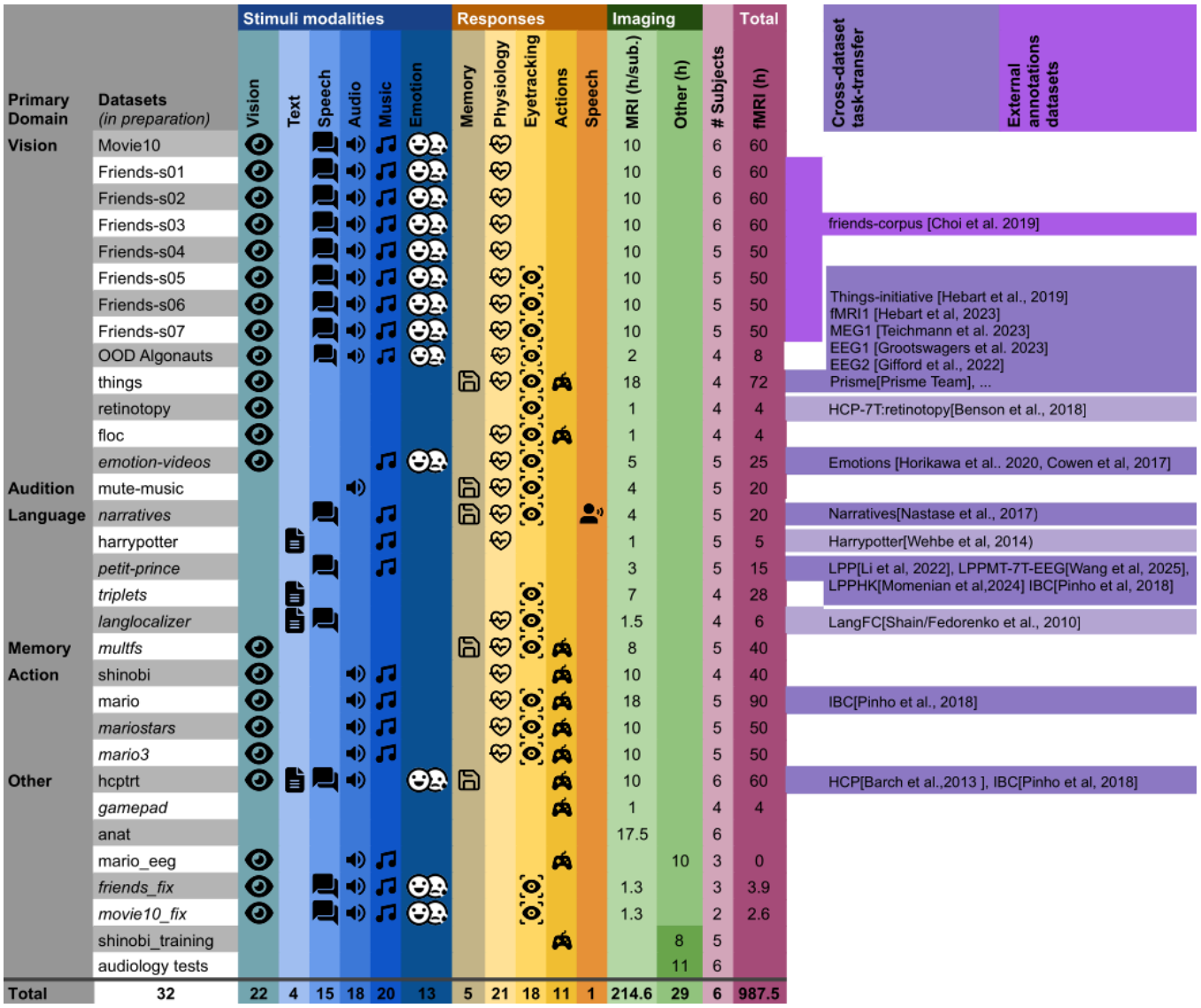
Data access & release

Raw and preprocessed fMRI data, behavioral responses and physiological recordings are formatted in BIDS (Gorgolewski, et al, 2016) and available via DataLad version control. After processing and quality checks, each dataset will be released independently with an accompanying data paper in the coming years. **Data for 4 subjects are accessible without any restrictions (CC0 license) on the Canadian Open Neuroscience Platform portal** (https://portal.conp.ca/), while data for all subjects is available via registered access at https://www.cneuromod.ca/.

Conclusion

The CNeuroMod project has assembled an unprecedented resource to model individual brain function using a wide range of controlled and naturalistic tasks. Data from the movie watching tasks are currently in use to assess the robustness of brain encoding models for the Algonauts 2025 challenge (Gifford et al., 2025). The wealth of additional data that will be released in the coming years will fuel novel insights into the ways human brains process complex stimuli.

Table 1. Summary of the datasets of the CNeuroMod project, detailing the tasks and cognitive domains probed, type and volume of collected data, as well as external datasets that allow studying generalizability across instruments and population sampling. Datasets names in *italics* are in preparations (i.e. scheduled for future release).



		Stimuli modalities						Responses					Imaging		Total		
Primary Domain	Datasets *(in preparation)*	Vision	Text	Speech	Audio	Music	Emotion	Memory	Physiology	Eyetracking	Actions	Speech	MRI (h/sub.)	Other (h)	# Subjects	fMRI (h)	Cross-dataset task-transfer / External annotations datasets
Vision	Movie10	●		●	●	●	●		●				10		6	60	
	Friends-s01	●		●	●	●	●		●				10		6	60	
	Friends-s02	●		●	●	●	●		●				10		6	60	
	Friends-s03	●		●	●	●	●		●				10		6	60	friends-corpus [Choi et al. 2019]
	Friends-s04	●		●	●	●	●		●				10		5	50	
	Friends-s05	●		●	●	●	●		●	●			10		5	50	
	Friends-s06	●		●	●	●	●		●	●			10		5	50	Things-initiative [Hebart et al., 2019]
	Friends-s07	●		●	●	●	●		●	●			10		5	50	fMRI1 [Hebart et al, 2023]
	OOD Algonauts	●		●	●	●	●		●	●			2		4	8	MEG1 [Teichmann et al. 2023]
	things	●						●	●	●	●		18		4	72	EEG1 [Grootswagers et al. 2023]
	retinotopy	●								●			1		4	4	EEG2 [Gifford et al., 2022]
	floc	●							●	●	●		1		4	4	Prisme[Prisme Team], ...
	emotion-videos	●				●	●		●	●			5		5	25	HCP-7T:retinotopy[Benson et al., 2018]
Audition	mute-music				●			●	●	●			4		5	20	Emotions [Horikawa et al.. 2020, Cowen et al, 2017]
Language	*narratives*			●		●		●	●	●		●	4		5	20	
	harrypotter		●			●			●				1		5	5	Narratives[Nastase et al., 2017)
	petit-prince			●		●							3		5	15	Harrypotter[Wehbe et al, 2014)
	triplets		●							●			7		4	28	LPP[Li et al, 2022], LPPMT-7T-EEG[Wang et al, 2025],
	langlocalizer		●	●					●	●			1.5		4	6	LPPHK[Momenian et al,2024] IBC[Pinho et al, 2018]
Memory	*multfs*	●						●	●	●	●		8		5	40	LangFC[Shain/Fedorenko et al., 2010]
Action	shinobi	●			●	●			●		●		10		4	40	
	mario	●			●	●			●	●	●		18		5	90	IBC[Pinho et al., 2018]
	mariostars	●			●	●			●	●	●		10		5	50	
	mario3	●			●	●			●	●	●		10		5	50	
Other	hcptrt	●	●	●	●		●	●			●		10		6	60	HCP[Barch et al.,2013], IBC[Pinho et al, 2018]
	gamepad										●		1		4	4	
	anat												17.5		6		
	mario_eeg	●			●	●					●			10	3	0	
	friends_fix	●		●	●	●	●			●			1.3		3	3.9	
	movie10_fix	●		●	●	●	●			●			1.3		2	2.6	
	shinobi_training										●			8	5		
	audiology tests													11	6		
Total	32	22	4	15	18	20	13	5	21	18	11	1	214.6	29	6	987.5	

Acknowledgements

The Courtois NeuroMod project was made possible by a grant from la Fondation Courtois given to LPB.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... & Kay, K. N. (2021). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nature Neuroscience, 24(7), 1166–1177. https://doi.org/10.1038/s41593-021-00962-x

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R. A., Smith, S., Johansen-Berg, H., Snyder, A. Z., & Van Essen, D. C. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage, 80*, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Benson, N. C., Jamison, K. W., Arcaro, M. J., Vu, A. T., Glasser, M. F., Coalson, T. S., Van Essen, D. C., Yacoub, E., Ugurbil, K., Winawer, J., & Kay, K. (2018). The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. *Journal of Vision, 18*(13), 23. https://doi.org/10.1167/18.13.23

Boudreau, M., Karakuzu, A., Boré, A., Pinsard, B., Zelenkovski, K., Alonso-Ortiz, E., ... & Cohen-Adad, J. (2025). Longitudinal reproducibility of brain and spinal cord quantitative MRI biomarkers. Imaging Neuroscience,3. https://doi.org/10.1162/imag_a_00409

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data, 6*, 49. https://doi.org/10.1038/s41597-019-00527-X

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings of the National Academy of Sciences, 114(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe, A., ... & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. Nature Methods, 16(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. Journal of Neurophysiology, 104(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Fortier, E., Bellec, P., Boyle, J. A., & Fuente, A. (2025). MRI noise and auditory health: Can one hundred scans be linked to hearing loss? The case of the Courtois NeuroMod project. PLOS ONE, 20(1), e0309513. https://doi.org/10.1371/journal.pone.0309513

Gifford, A. T., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage, 264*, 119754. https://doi.org/10.1016/j.neuroimage.2022.119754

Gifford, T., Bersch, D., St-Laurent, M., Pinsard, B., Boyle, J., Bellec, L., Oliva, A., Roig, G., & Cichy, R. M. (2025). arXiv preprint arXiv:2501.00504.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., ... Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data, 3,160044.https://doi.org/10.1038/sdata.2016.44

Grootswagers, T., Zhou, I., Robinson, A. K., Henderson, R., Cichy, R. M., & Carlson, T. A. (2022). *Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams*. Scientific Data, 9, 3. https://doi.org/10.1038/s41597-021-01102-7

Harel, Y., Cyr, A., Boyle, J., Pinsard, B., Bernard, J., Fourcade, M.-F., ... & Fortier, E. (2023). Open design of a reproducible videogame controller for MRI and MEG. PLOS ONE, 18(11), e0290158.

Hebart, M. N., Zhang, C. Y., Pereira, F., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. PLOS ONE, 14(10), e0223792. https://doi.org/10.1371/journal.pone.0223792

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., & Baker, C. I. (2023). THINGS-data, a multimodal

collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife, 12*, e82580. https://doi.org/10.7554/eLife.82580

Horikawa, T., Cowen, A. S., Keltner, D., & Kamitani, Y. (2020). The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. iScience, 23(5), 101060. https://doi.org/10.1016/j.isci.2020.101060

Kay, K. N., Winawer, J., Mezer, A., & Wandell, B. A. (2013). Compressive spatial summation in human visual cortex. *Journal of Neurophysiology, 110*(2), 481–494. https://doi.org/10.1152/jn.00105.2013

Li, J., Bhattasali, S., Zhang, S., Lal, N., Fedorenko, E., & Brennan, J. R. (2022). Le Petit Prince multilingual naturalistic fMRI corpus. Scientific Data, 9, 530. https://doi.org/10.1038/s41597-022-01625-7

Ma, K., Jurczyk, T., & Choi, J. D. (2018, June). Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 2039-2048). https://doi.org/10.18653/v1/N18-1185

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., ... & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. Nature Neuroscience, 25, 1014–1019. https://doi.org/10.1038/s41593-022-01114-5

Momenian, M., Ma, Z., Wu, S., Chan, Q., Chan, C. C. H., & Wong, A. (2024). Le Petit Prince Hong Kong (LPPHK): Naturalistic fMRI and EEG data from older Cantonese speakers. Scientific Data, 11, 992. https://doi.org/10.1038/s41597-024-03745-8

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., ... & Hasson, U. (2021). The “Narratives” fMRI dataset for evaluating models of naturalistic language comprehension. Scientific Data, 8, 250. https://doi.org/10.1038/s41597-021-01033-3

Pinho, A., Amadon, A., Ruest, T., et al. (2018). Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping. Scientific Data, 5, 180105. https://doi.org/10.1038/sdata.2018.105

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. Cognitive Neuroscience, 8(3), 167–176. https://doi.org/10.1080/17588928.2016.1201466

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., & van Gerven, M. A. J. (2019). A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. bioRxiv. https://doi.org/10.1101/687681

Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal processing capacity in high-level visual cortex is domain specific. The Journal of Neuroscience, 35(36), 12412–12424. https://doi.org/10.1523/JNEUROSCI.4822-14.2015

Wang, Q., Zhou, Q., Ma, Z., Wang, N., Zhang, T., Fu, Y., & Li, J. (2025). Le Petit Prince (LPP) Multi-talker: Naturalistic 7T fMRI and EEG dataset.*bioRxiv*. https://doi.org/10.1101/2025.02.26.640265

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, 10(3), e0123148. https://doi.org/10.1371/journal.pone.0123148