**基于大语言模型表征一致性的情绪识别机制研究**

**项目背景：**

情感（Emotion）在生物智能发展和学习中的重要性已得到广泛验证，它深刻影响着生物体的行为和决策(Damasio, 2006; LeDoux, 2000; Tyng et al., 2017)。近年来，随着人工智能技术的迅速发展，研究者开始探讨情绪与情感在人工智能系统中的应用，旨在提升模型的情感识别、理解和生成能力，从而实现更加自然和人性化的人机交互(Iqbal et al., 2018; Li et al., 2018)。目前，经过大量人类自然语言文本预训练的大语言模型（Large language model, LLM）已具备一定的人类情感概念知识，如BERT系列，GPT系列，Deepseek等模型，​这使得它们在情感分析、情感生成等任务中表现出较好的性能(Wang et al., 2023)。然而，情绪本身作为人类的复杂心理现象，并不存在客观的实体作为标签，难以获得情绪识别任务的真实值（ground truth），这给LLM在情感领域的真实能力评估带来了挑战。因此，深入探究LLM在情绪识别任务中的机制，明确LLM在多大程度上捕获了与人类相似的情绪表征结构，对于提高人工神经网络（Artificial Neural Network, ANN）在情绪处理任务上的表现至关重要。

已有研究证实，人工神经网络的表征与人类大脑的表征存在一定相似性，从最早的视觉领域(Eickenberg et al., 2017; Yamins et al., 2014)，到语音处理(Kell et al., 2018; Koumura et al., 2023)以及语言理解(Abdel-Ghaffar et al., 2024; Mischler et al., 2024; Schrimpf et al., 2021)，并且对于在模型架构、训练方式以及训练数据等方面存在差异的ANN表征均能较好地预测人类大脑表征。例如，最近在视觉领域中的研究表明不同视觉神经网络之间以及神经网络与人类视觉皮层之间存在共同的表征特征维度(Chen & Bonner, 2024)；而在另一项研究中，研究者分别研究了语言和视觉的跨模型表征一致性，结果发现输入刺激在不同ANN的高维嵌入中表征相似性越高（表征越一致），模型表征与人脑对该刺激的表征也会越相似（图1）(Hosseini et al., 2024)。这些结果提示，在经过大规模且多样化的视觉或语料数据的学习后，不同模型学习到了共同的隐含世界知识（表征），而不是特定模型的特征(Huh et al., 2024)。这不仅揭示了模型在学习过程中对世界知识的归纳能力(Conwell et al., 2024)，还为研究人类的语言与认知提供了新的视角，帮助探索语言理解、概念形成及知识共享的机制。

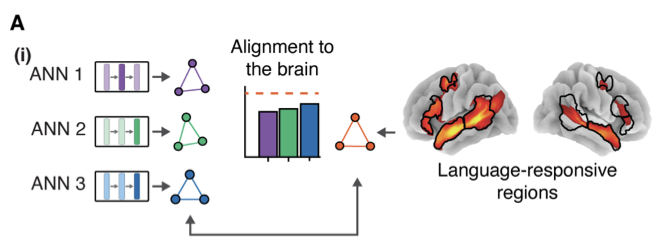


图 1 高一致性表征的刺激在ANNs和人脑间收敛

目前高性能大语言模型在情感任务上的表现，离不开预训练过程中对人类语言文本的大量学习，情感建构理论表明，语言经验有助于构建离散的情感概念(Barrett, 2017; Barrett et al., 2007)，而情感概念的表征对于情绪认知来说十分重要(Winkielman et al., 2018)。从情绪概念发展的角度来看，儿童随着语言输入变得越来越多、越来越有用，情绪词可能会成为一种线索，将共享较少感知特征的事例们整合在一起，并施加功能上的相似性，帮助儿童修正其内部模型，使之与其文化中的成人模型相似(Hoemann et al., 2019)。在情感计算领域，也有研究发现神经网络神经元基于语言的情感表征对推断新情境下情感含义的能力有因果关系，进一步表明语言在构建和组织情绪概念中的核心作用(Li et al., 2024)。因此，如果LLM在大规模文本数据预训练后，能够自发形成稳定且跨模型通用的情绪概念表征，则可能说明语言内部本身具有稳固的情绪概念结构。具体来说，它们在同样的语言文本刺激中提取到了相同的情绪概念，从而能够提升下游任务的表现，并且在此基础上神经网络对文本的表征向量如果能进一步预测人类大脑对该文本的响应，这将进一步证明 LLMs 的语义表征机制与人类情绪认知存在潜在的对应关系。

综上所述，本研究推测在情绪任务中探索不同LLM之间情绪表征一致的情绪概念及相关属性，将有助于在复杂的情绪概念表征中找到一种普遍适用的表征结构，从而理解人工智能系统情绪识别的机制，并为开发更具泛化性与可解释性的情绪人工智能系统提供重要指导。随后，为进一步验证这些表征的心理可解释性，将对比模型在不同一致度表征下对人脑情绪反应的预测能力，和相关脑区对情绪属性解码的能力，以探究这些共通表征是否映射了人类的情绪加工方式。

**研究问题与方案：**

**研究一：基于大语言模型之间的表征相似性，探究神经网络情绪类别概念的表征**

在这一部分的研究中，本课题将探究不同高性能LLM对情绪类别概念的表征一致性，探究其与人类情绪认知的相似性。具体来说，为了把情绪概念表征从句子语义表征中分离出来，本研究参考Li等人2024年的工作，采用Prompt tuning的方法，当大语言模型完成文本情感分类任务时，为每个情绪类别的句子添加一段特定于该类别的prompt（图2A），这个在优化过程中可学习的prompt经过LLM编码后的隐向量被认为是该情绪类别概念的表征（图2B）。这一方法的优势在于概念表征的过程中没有输入样本文本，任务抽象方法可防止知识表征可能受到刺激关键词的污染。

接下来，采用Hosseini（2024）等人的方法，在该任务的情绪类别中基于表征相似性的方法，通过计算类别刺激的模型间相关性，得到跨模型间表征一致性高和低的两组情绪类别集合。根据这篇研究的发现，表征一致性相似的文本刺激在生活中出现的频率更高，并且更为具象，因此，在情绪类别的概念表征中，本研究假设基本情绪类别在跨模型间表征更一致，复合情绪的表征一致性更低。在此基础之上，将分别测试当一致性表征高和低类别集合中的句子作为情绪文本材料时，LLM在情绪识别任务上的表现，假设跨模型的一致性表征代表不同模型共同学习到了情绪概念中稳固的结构，因此相比于低一致性表征的情绪类别，在高一致性表征集合中，不同类别的区分度更高，进而情绪识别任务的准确率更高。

**高性能LLM模型** 根据前人研究(Hosseini et al., 2024; Schrimpf et al., 2021)，本研究选择以下7个大语言模型，它们不仅在语言理解和生成具备一定的能力，并且对语言网络中的 fMRI 体素级反应具有最佳的预测能力：RoBERTa: roberta-base (Liu et al., 2019), BERT: bert-large-uncased-whole-word-masking (Devlin et al., 2019), ALBERT: albert-xxlarge-v2 (Lan et al., 2019), XLM: xlm-mlm-en-2048 (Lample & Conneau, 2019), XLNET: xlnet-large- cased (Yang et al., 2019), CTRL: ctrl (Keskar et al., 2019), GPT2: gpt2-xl (Radford et al., 2019)。

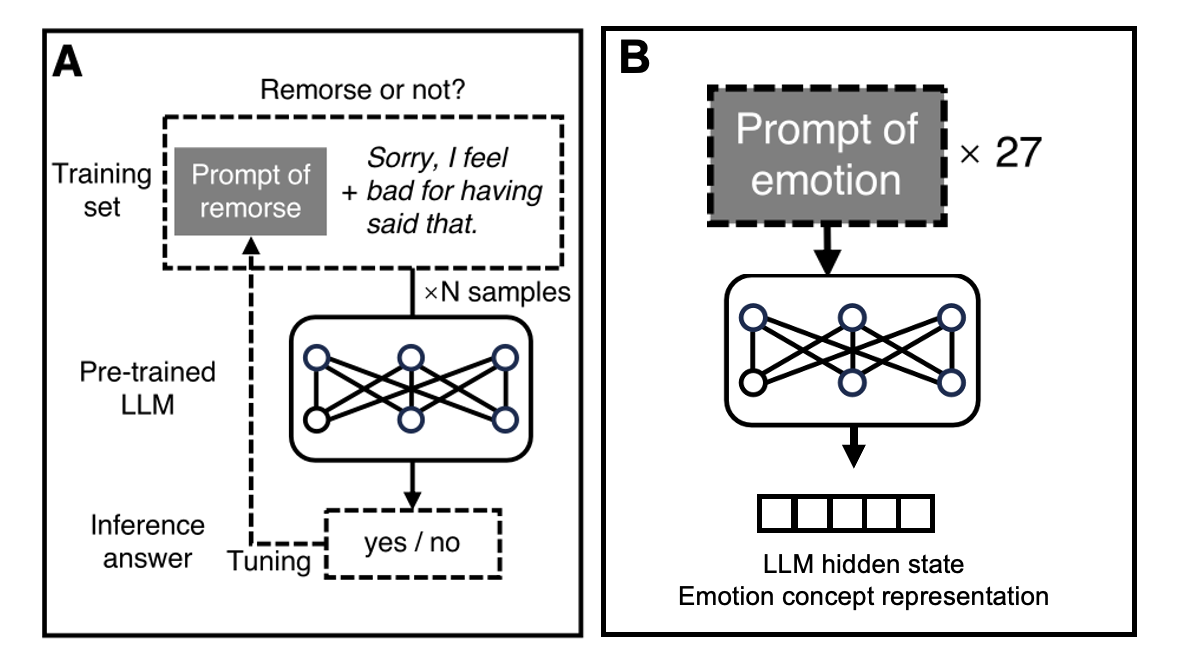


图 2 情绪概念的表征

**研究二：LLM情绪识别中跨模型表征一致性的可解释性机制研究**

在研究二中，本课题进一步探索研究大语言模型基于自然语言文本的情感分析任务（如情绪识别）中表征的可解释性机制。根据研究一中跨模型间表征一致性高和低两类情绪类别，进而得到类别中“表征高一致”与“表征低一致”两类文本刺激。首先，本研究探讨具体哪些情感概念属性（如效价、唤醒度、控制感、确定性、普遍性、具象性、语境完整性、文化特异性等）与大语言模型间的情绪语义表征一致性存在关联。具体而言，为了获得每个刺激的多维情感属性评分，将招募被试对所使用的情感刺激进行属性标注；随后，通过相关性分析、回归分析、机器学习等方法，检验这些属性评分与LLM之间表征一致性的关系。此外，本研究将继续探究跨模型表征一致性高、低两个刺激集中句子在情绪类别与情绪维度在文本上的映射关系，揭示大语言模型情绪概念组织结构的可解释性。

为更深入地研究LLM表征与人类大脑表征之间的关系，本研究计划收集8个被试在功能性核磁共振成像（functional magnetic resonance imaging, fMRI）中完成上述相同的情感任务，选择高、低一致性的刺激集作为实验材料，记录他们在任务过程中阅读文本材料时的神经响应。如果多个大语言模型之间的情绪语义表征一致性高，则预计相比于低一致性的刺激，不同大语言模型的表征均能更有效地预测人脑的神经响应（具体模型-大脑相似性计算方法参考Schrimpf, 2021）。

接下来，将继续探索LLM模型的表征在全脑的预测能力，并通过模型拟合指标找到显著预测的脑区，如果被LLM表征向量显著预测的脑区的体素响应能够解码与表征一致性相关联的情绪属性，则说明LLM与大脑对文本的编码收敛到同一情绪属性表征轴。研究二方案示意图见图3。

**任务数据库** 英文：GoEmotions (Demszky et al., 2020)；中文：<https://aistudio.baidu.com/competition/detail/50/0/task-definition>（需要对类别进行人工标注）。

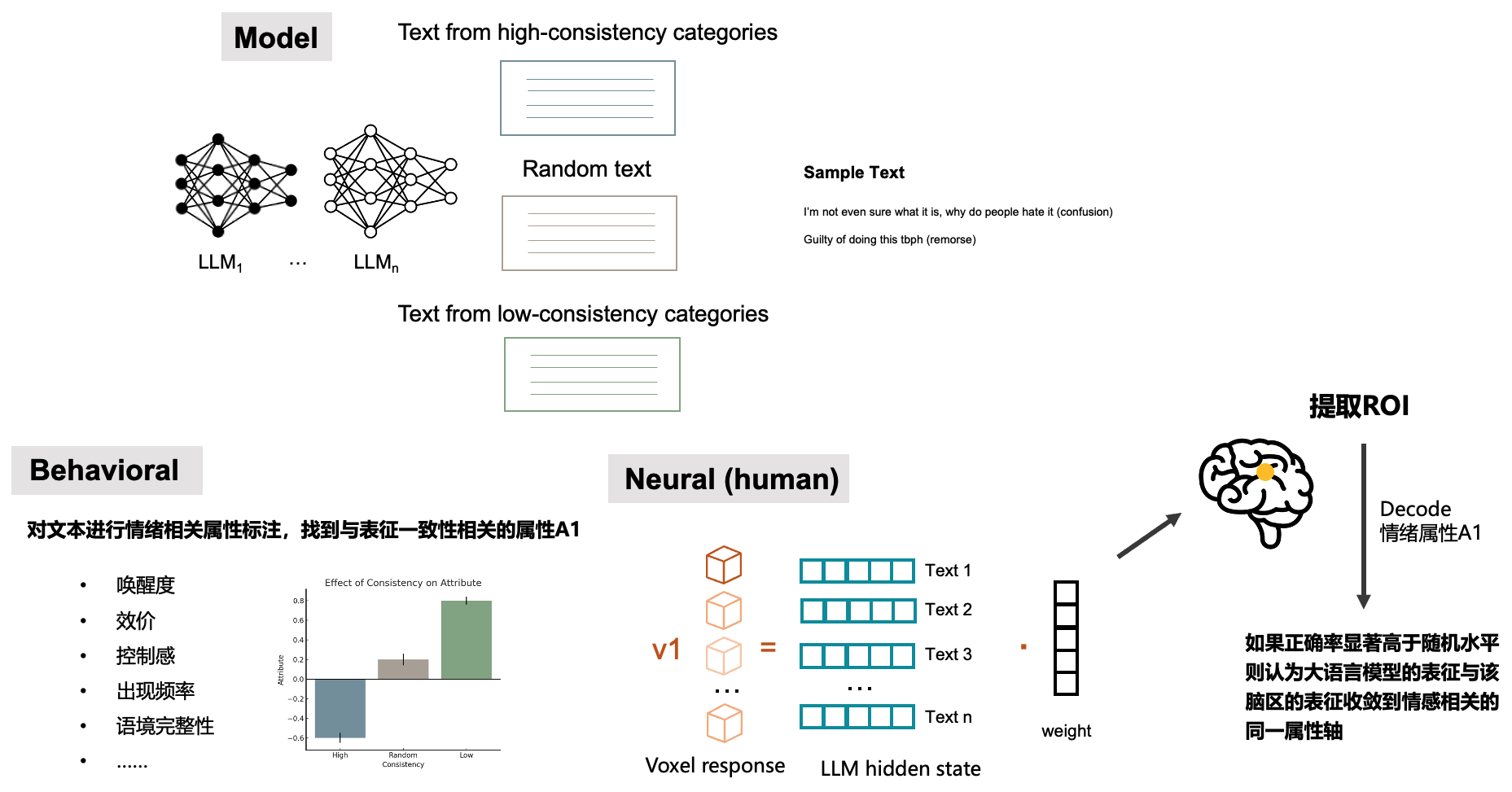


图 3 研究二方案示意图

**理论与实际意义：**

首先在机制和理论层面，本研究基于多个高性能大语言模型在情绪概念和语义上的表征相似性分析，探究学习了大量人类自然语言文本的人工智能神经网络是否学习到了共同的隐含世界知识，以及这一现象是否表明LLM在情感任务上的优异表现，通过对刺激的多个情感相关的属性进行分析，可以帮助我们了解跨模型情绪概念表征一致的可解释性机制，揭示LLM在情绪任务上的表现与其情绪表征结构之间的关系。

为进一步验证这些表征是否具有心理可解释性，当模型-模型之间表征一致性与性能关联被证实后，如果能进一步说明这些一致度较高的表征也能更好地预测人脑对情绪刺激的反应，则可能表明这类一致的表征结构并非偶然，而具有与人类类似的情绪加工‘原则’或‘映射’。这不仅能在一定程度上揭示人类自然语言与情绪知识概念表征的关系，同时也是对证明建构主义理论在人工智能语境下的适用性的印证。

在应用意义上，本研究可以定向目前在大语言模型的情绪概念表征和情绪任务中仍然欠缺的细颗粒度情绪类别，优化情感任务训练数据，使其在关键属性上具备足够的信息量，增强数据质量，实现仅用少量高质量的样本提升LLM情绪理解及情感分析能力。

**参考文献**

Abdel-Ghaffar, S. A., Huth, A. G., Lescroart, M. D., Stansbury, D., Gallant, J. L., & Bishop, S. J. (2024). Occipital-temporal cortical tuning to semantic and affective features of natural images predicts associated behavioral responses. *Nature Communications*, *15*(1), 5531.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, *12*(1), 1-23.

Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in cognitive sciences*, *11*(8), 327-332.

Chen, Z., & Bonner, M. F. (2024). Universal dimensions of visual representation. *arXiv preprint arXiv:2408.12804*.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, *15*(1), 9383.

Damasio, A. R. (2006). *Descartes' error*. Random House.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers),

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184-194.

Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental psychology*, *55*(9), 1830.

Hosseini, E., Casto, C., Zaslavsky, N., Conwell, C., Richardson, M., & Fedorenko, E. (2024). Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024.2012. 2026.629294.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Iqbal, M. T. B., Abdullah-Al-Wadud, M., Ryu, B., Makhmudkhujaev, F., & Chae, O. (2018). Facial expression recognition with neighborhood-aware edge directional pattern (NEDP). *IEEE Transactions on Affective Computing*, *11*(1), 125-137.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630-644. e616.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Koumura, T., Terashima, H., & Furukawa, S. (2023). Human-like modulation sensitivity emerging through optimization to natural sound recognition. *Journal of Neuroscience*, *43*(21), 3876-3894.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual review of neuroscience*, *23*(1), 155-184.

Li, M., Su, Y., Huang, H.-Y., Cheng, J., Hu, X., Zhang, X., Wang, H., Qin, Y., Wang, X., & Lindquist, K. A. (2024). Language-specific representation of emotion-concept knowledge causally supports emotion inference. *iScience*, *27*(12).

Li, Y., Zeng, J., Shan, S., & Chen, X. (2018). Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, *28*(5), 2439-2450.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., & Mesgarani, N. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, 1-11.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Tyng, C. M., Amin, H. U., Saad, M. N., & Malik, A. S. (2017). The influences of emotion on learning and memory. *Frontiers in psychology*, *8*, 235933.

Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, *17*, 18344909231213958.

Winkielman, P., Coulson, S., & Niedenthal, P. (2018). Dynamic grounding of emotion concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170127.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619-8624.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.