

Perspective

Data science opportunities of large language models for neuroscience and biomedicine

Danilo Bzdok,^{1,3,*} Andrew Thieme,² Oleksiy Levkovskyy,² Paul Wren,² Thomas Ray,² and Siva Reddy^{1,4,5}

¹Mila - Quebec Artificial Intelligence Institute, Montreal, QC, Canada

²Mindstate Design Labs, San Francisco, CA, USA

³TheNeuro - Montreal Neurological Institute (MNI), Department of Biomedical Engineering, McGill University, Montreal, QC, Canada

⁴Facebook CIFAR AI Chair

⁵ServiceNow Research

*Correspondence: danilobzdok@gmail.com

<https://doi.org/10.1016/j.neuron.2024.01.016>

Large language models (LLMs) are a new asset class in the machine-learning landscape. Here we offer a primer on defining properties of these modeling techniques. We then reflect on new modes of investigation in which LLMs can be used to reframe classic neuroscience questions to deliver fresh answers. We reason that LLMs have the potential to (1) enrich neuroscience datasets by adding valuable meta-information, such as advanced text sentiment, (2) summarize vast information sources to overcome divides between siloed neuroscience communities, (3) enable previously unthinkable fusion of disparate information sources relevant to the brain, (4) help deconvolve which cognitive concepts most usefully grasp phenomena in the brain, and much more.

INTRODUCTION

Language has more human information per bit than potentially any other form of data. Natural language processing (NLP) to analyze human text has come a long way. In the early days, simple language models like n-gram models (e.g., a 2-gram treats word-word combinations as unique entities) were used to study language and semantics with various goals. Language models have at times also been used to study various cognitive tasks like reading comprehension, language translation, and question answering. Researchers compared the performance of NLP models on these tasks with human performance to gain insights into human cognition, such as in the field of psycholinguistics.

The rise of deep learning after ~2010 ignited the era of semantic “embeddings” in NLP modeling: single words, sentences, paragraphs, or entire documents could be encapsulated in a compact float-vector format that denotes meaning. Intuitively, such embeddings can be thought of as locations in a high-dimensional coordinate system that enable mapping of semantic entities (word sequences) relative to their contextual similarity.^{1–4} The more two semantic entities denote similar contexts, the more similar their semantic embeddings will be. Using last-generation models like Word2Vec⁵ and GloVe,⁶ researchers started to use these interoperable semantic embedding representations to quantify the relationships in meaning, such as between words or sentences.

Current LLMs are trained on more text than one human could read in hundreds or thousands of lifetimes. This allows them to perform feats like writing computer programming code, mathematics, planning, literature reviews and summations, or playing text-based games—many forms of emergent capabilities that even their developers did not anticipate.⁷ Sometimes such models are used to study how the brain itself processes contextual information and how the human mind generates language

(please see Goldstein et al.,⁸ Caucheteux et al.,⁹ and Schrimpf et al.¹⁰ for excellent examples), although this will not concern us here. As the current paradigm shifts and trends exponentially, LLMs learn what are probably the most powerful internal representations of meaning to date.

Human language mirrors human thought, which is why state-of-the-art NLP is likely to provide us with inherent advantages. In this perspective, we attempt to discuss impending implications for investigators in neuroscience and biomedicine.

DATA SCIENCE PERSPECTIVE ON LARGE LANGUAGE MODEL SOLUTIONS

Historically, convolutional deep neural networks have revived AI excitement since 2010–2012, and LLMs are currently fueling yet another wave of momentum in the AI ecosystems. Language modeling has recently made substantial leaps forward following the introduction of the transformer architecture (for example, Vaswani et al.¹¹ was cited >90,000 times in the first 5 years after publication), driving the current thrust in AI innovation. For example, GPT-2 consists of 24 transformer blocks and recent architectures are even deeper (some of which have not been disclosed). As an instance of what has recently been called “generative AI” (gen AI), the outcome from these algorithms is not a class (e.g., patients with disease versus healthy group), number (e.g., cognitive performance measure), or discrete category (e.g., brackets of yearly income), but a structured “content” like language (as well as images or audio formatted information), i.e., synthesizing or fantasizing new content from previously ingested content.

Doing away with much of the complexity in the previous deep-neural-network generation, transformers have become state-of-the-art in NLP (Figures 1 and 2). This simplified architecture is



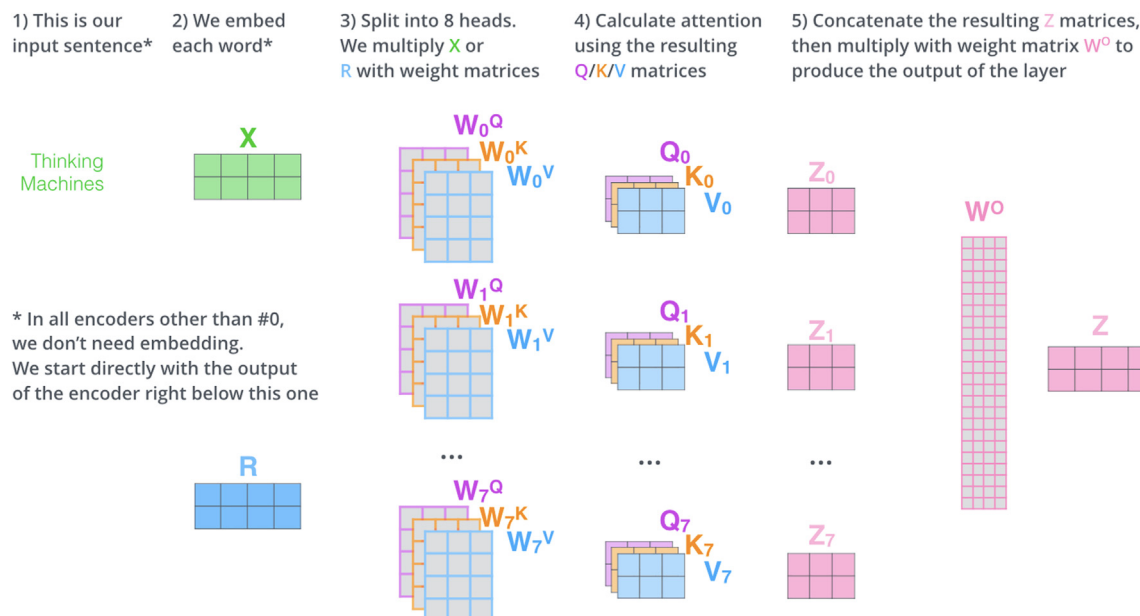


Figure 1. The heart of the self-attention mechanism in transformer architectures like LLMs

A sentence of two exemplary words (tokens) are re-represented as a sequence of two semantic embeddings (X , 4-dimensional semantic embedding), one row for each input word occurrence. Then, based on the two-word embedding from the encoder neural network (cf. Figure 2), self-attention invokes three new vectors: query (Q), key (K), and value (V). Each of these three vectors emerges from a matrix multiplication between X and a to-be-estimated matrix W_0 , whose parameter entries are trained in conjunction with the overall neural network. The ensuing query vector, key vector, and value projection vectors typically have smaller dimensionality than the embedding for each word of the input sequence. Regarding the content of representation, query vectors instantiate a focus subset of the input sequence (“the question we are asking”), key vectors instantiate the entire input sequence (“responses for everything we could ask”), and value vectors instantiate a corresponding quantity (“the content of all possible answers”). Using the queries, keys, and values, the attention mechanism creates weighted attention scores as output derived from X (“how much attention to pay to each response based on relevance to the question”). In the transformer model, this attention mechanism is used multiple times in each layer and across multiple layers to progressively refine the representations of the input sequence (Figure 2) (source: <https://jalammar.github.io/illustrated-transformer/>).

much more scalable than its predecessors, partly because this modeling setup also lends itself well to parallelization of computation workflows. In contrast to previous deep NLP solutions, in transformer architectures, the interdependencies between word tokens, close or far, are equally well captured. Also departing from certain previous neural network design, transformer models are feed-forward deep learning architectures that do not include explicit loops of processing. Instead, there are implicit loops created by contextualization of the already generated, previous text fed back into the LLM as input (“autoregression”). Unlike the previous LLM generations revolving around BERT (bidirectional encoder representations from transformers; using a word-masking logic as the modeling objective), generative pre-trained transformer (GPT) architectures, such as ChatGPT, allocate attention only to the left-hand reading context or previous word tokens during training, which results in its unidirectional mode of processing, specifically autoregressive in nature. So-called position encoding is a feature of these architectures that quantifies a notion of where a word appears in a sentence (not just an unsorted “bag of words”). In other words, GPT variants predict the next word token in a sequence based on the preceding word tokens. Because of its unidirectional nature, GPT LLMs do not “see” or “consider” subsequent tokens when predicting the next word—it is looking to the past, not the future, of a given sentence (with analogy to how a human reads a book).

It is the self-attention mechanism that is at the heart of the transformer-augmented and GPT-like modeling architectures. This feature allows the model to allocate sensibility to different segments of the input text sequence based on different “attention scores.” Dedicating attention to closer versus further away word tokens turns out to be algorithmically identical—no iterative, step-by-step process is needed to involve information segments further away, as is required in earlier generations of deep learning architectures. In transformers, the mechanism of focusing on nearby or distant words in a sentence is handled in the same way, allowing the model to consider all parts of a sentence or text sequence simultaneously. In contrast to earlier neural networks, this does not require a sequential, “time-step” approach to process and coagulate distant ends of the input. The computational complexity of the common implementation of the self-attention mechanism scales quadratically with sequence length.¹² Despite various improvements in attention mechanisms, most of them still struggle in certain use cases involving especially long sequences.¹³ Each transformer layer can “see” all tokens in its scope at once. Nevertheless, the depth of recursive information processing is limited by the number of consecutive transformer layers, such as in levels of nested meaning of sentences or nested multiplication of number sequences.

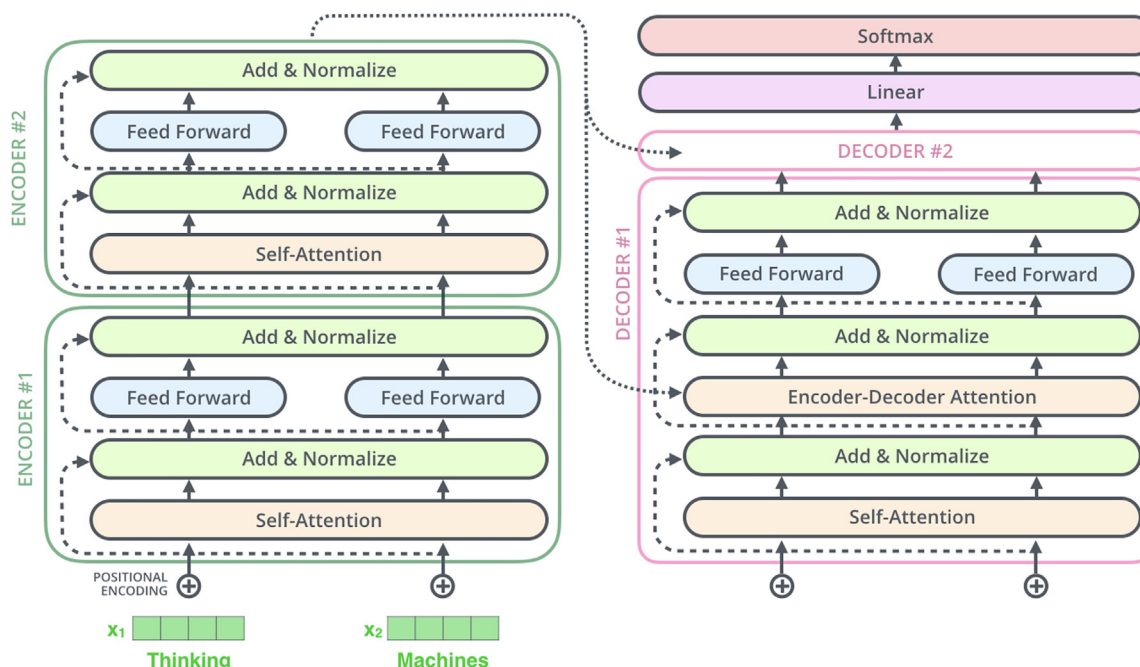


Figure 2. The role of the self-attention layer in the transformer neural network architecture

Each self-attention layer is followed by a normalization layer. After the self-attention mechanism, each word vector is individually passed through a feed-forward neural network. The same feed-forward network is used for each word, but the parameters are not shared between different positions in the sentence (this operation is applied independently to each position). The output of one layer is fed into the consecutive neural network layer. Overall, the encoder block (left) converts ambient input data (e.g., sequence of words) into a structured hidden representation in the form of semantic embedding vectors of that input sequence. The decoder block (right) uses the result of the encoder layers to go from that latent semantic representation to continuous-value vectors that are finally transformed into raw output data (e.g., word sequence in a different language), in the form of probability distribution across the vocabulary (softmax, red). Each probability value in the softmax output indexes a unique token in the vocabulary. Concretely, the highest-probability token (e.g., output word in next-word prediction in a sentence) gets picked as output product of the current processing step (source: <https://jalammar.github.io/illustrated-transformer/>).

Moreover, current LLM architectures typically have several parallel attention mechanisms lined up in each of the consecutive transformer layers. This “multi-headed attention” (1) enables placement of a simultaneous parallel focus on several different aspects of the input sequence, expanding the breadth of complexities that can be captured overall, and (2) thus allows for several dimensions of semantic representation to be identified and extracted all at once (with some analogy to modeling different latent factor components like principal component analysis or autoencoder neural networks).¹⁴

A notable and practically relevant choice is the temperature parameter (a scalar value in the range: 0 to positive number). This hyperparameter controls the degree of “creativity” in the model outputs, as a form of calibrating exploration versus exploitation. Setting high temperature (e.g., >1) produces a softer probability distribution in the word nomination in the last model layer. This leads to intentionally more fuzzy, and thus potentially less accurate but also more creative, outputs. In contrast, low temperature (e.g., <1) leads to a sharper probability distribution in output word relevances. In this mode of operation, the model becomes more deterministic, sticking closely to the most probable candidates in the output distribution, thereby reducing stochasticity in its responses.

Despite simple modeling objectives (e.g., BERT invoking word masking and GPT3 invoking next-word prediction, while

involving human feedback in the case of GPT4/ChatGPT), by their sheer enormity, transformer-endowed architectures have sparked few-shot learning (i.e., learning based on few examples from a target task) and performance paradigms, inherently deriving aspects of a semantic world model in multiple settings.¹⁵ These capabilities are at the core of the self-supervised modeling regime (cf. next sections). These secondary consequences led even the creators of these models to struggle for explanations behind the successes of LLMs.¹⁶

EMERGING SCALING LAWS OF LARGE LANGUAGE MODEL SOLUTIONS

What are the limits of scale? As a key quality for impact, LLMs yield a rapidly higher quality of model instance with increasing number of training observations. Having roughly 2–20 times more training word tokens than model parameters for model fitting has led to impressive performances on multiple occasions. From a data perspective, it is challenging to get a sense of the upper bound on available text, text-transformed, and text-transformable data. As one concrete consideration, the size of the entire text volume on the internet may today reach around 2 trillion word tokens, based on simple normative assumptions (1.2 billion websites $\times$ 1,500 words per website on average [according to ChatGPT query, September 2023]). From a model perspective, from 2018

to 2022, the sizes of the LLMs have increased from $\sim 10^8$ (e.g., ELMo, BERT-L) to $\sim 10^{11}$ (e.g., PaLM) parameters to be estimated. As a first rule of thumb holding across many goals and applications, expanding the depth and width of the model (increasing the number of parameters) leads to clear performance improvements. Knowing how a model scales is of strategic value as such insights inform decisions on resource allocation: how to prioritize compute budget, data troves, and model size.

More specifically, one comprehensive, widely regarded empirical study in the deep learning literature explored and carefully benchmarked seven orders of magnitude of model scale.¹⁷ These investigators designed computational experiments that successfully converged on three key factors that determine model scaling: (1) the number of model parameters (N), (2) the amount of available data (D), and (3) the amount of available computation power (C) used for model estimation. In stark contrast, in these experiments, the model performance only mildly depended on the actual shape of the model architecture. Overfitting (i.e., over-adjustment to idiosyncrasies in the training data) appeared to be largely prevented by increasing N and D in parallel. In contrast, performance decay resulted if only N or only D was increased (but see Touvron et al.¹⁸), holding the respective other factor fixed. Finally, continued scale up of N, D, and C displayed patterns of diminishing returns, following a power law.

However, as a very recent development, rather than continuing the initial trend of increasing model size, LLMs have been shrunk back more and more in terms of parameter number.^{18,19} Counterintuitively to many investigators, reducing the model size again, potentially aligning better with the actual amount of available data, boosted the model performance, loosened up memory requirements, and relieved the computational cost. These improvements may turn out to be critical for the applicability of LLM solutions to real-world problems and increase the potential, for example, of smart phones carrying dedicated LLMs in the years to come. In short, a nascent research stream indicates that more data are, relatively, more important than larger model sizes in terms of parameters, while both are driving factors, each by itself.

Of note, measuring model performance critically depends on the investigator's choice of evaluation metric.²⁰ These authors argue that changes in LLM behavior deemed "emergent" (abilities that are not readily apparent in smaller-scale models but are present in large-scale models¹⁵) may become apparent only due to researchers' choice of certain evaluation metrics. Conversely, the authors²⁰ also showed that metric choice can induce seeming emergent abilities in diverse architectures and tasks. Hence, recent empirical investigations show²⁰ that changing metrics can weaken or strengthen signs of emergent abilities in LLM architectures as a function of model scale, with direct implications for AI safety and AI alignment.

Overall, larger LLMs turn out to be more sample efficient than smaller LLMs in the fine-tuning or few-shot learning scenario. That is, paradoxically, the more model parameters need to be estimated, the less input data points you need to achieve comparable performance. As in data science in general, increased data quality can always lead to further performance gains. Although it is important to acknowledge that neural network scaling laws are still almost entirely empirical at this point, these

scaling behaviors show robust trends (but see Caballero et al.²¹). The expansion and explosion of LLM architectures was fueled by (1) the invention of transformers, which tend to vary only slightly across recent LLMs, (2) availability of abundant data sources, and (3) availability of compute power at scale. Of relevance to the next section, the specific architecture of the model (number of layers, layer dimension, etc.) is relatively inconsequential, particularly as the model size increases.

LARGE LANGUAGE MODELS EXHIBIT UNPRECEDENTED TRANSFER LEARNING CAPABILITIES

For deep learning tools to thrive, there is commonly a need for data abundance. However, many areas of neuroscience do not have massive data troves readily available, let alone the internet-scale kinds of datasets that fuel text and image analysis in the AI community. This discrepancy begs the tactical question: what kinds of abundant non-neuroscience data can be leveraged to port modeling solutions over to revisit and attack neuroscience problems?

Intuitively, "transfer learning" is a mode of data analytics that revolves around storing structured knowledge gained while solving one problem, for applying it to a different but related problem. Transfer learning aims to improve the performance on a similar, often more constrained, modeling task that is typically (severely) under-resourced. In the context of deep learning, this modeling regime typically refers to the practice of, as a starting point, pre-training a model on a massive dataset and, then, applying or refining ("fine-tuning" its model parameters by slight adjustments) this model on a smaller dataset pertaining to a specific task of actual interest (please see <https://www.ruder.io/transfer-learning/> for a comprehensive source on fine-tuning techniques for LLMs). This agenda cashes in on the hypothesis that the features learned by the pre-trained model can serve as a general representation, beneficial for the target task. Historically, the success of transfer learning typically depended on a high degree of similarity between the pre-training and fine-tuning tasks (but see next section).

LLMs, and other transformer-carrying architectures, have shown beyond-expectations capability in transfer learning, thus revolutionizing NLP by expanding the scope of executable tasks. As a key inflection point, until recently, the dominant paradigm still consisted in *supervised* model pre-training on massive corpora. This requirement of large quantities of data points with high-quality annotations was vexing. High-quality labels are typically logistically challenging to obtain—severely limiting what kinds of data available on the internet and other sources could actually be used for effective pre-training and thus transfer learning. It is only now that *unsupervised* pre-training, not requiring accurate annotations for each data point, came into reach and generated unseen performances. This watershed event considerably expands the scope of data usable for pre-training of LLMs.

More formally, the more parameters that need to be estimated in an LLM, the slower the model development process. LLMs opened the door to new regimes of fine-tuning going beyond what pattern-learning algorithms could achieve before. Several

approaches have been proposed to adapt a model to a new task while only updating or adding a relatively small number of parameters. One tactic consists of “freezing” (leaving unchanged) the parameters of several layers of a pre-trained LLM. This approach then adapts only a small fraction of adjustable parameters for the downstream task, thus avoiding “(catastrophic) forgetting” of originally extracted knowledge encapsulated in the initial model instance.

This logic can be extended during fine-tuning by adding new learnable layers within the LLM. Such “adapters”^{22–24} can considerably reduce training time and compute costs on the target task. The selection of particularly high-quality data for the fine-tuning phase was shown to lead to competitive performance, with even fewer target task examples. LLMs proved remarkable in few-shot learning. At its extreme, zero-shot learning leveraging pre-trained LLMs (using a trained LLM on a new task, without providing examples for that new task) turned out to be proficient at a variety of downstream tasks out-of-the-box, that is, even without adjustment of the pre-trained model.^{25,26}

In short, the monstrosity of LLMs, encompassing billions of adjustable model parameters, unlocked the extraction of quintessential representations from massive text corpora, without the previously acute need for supervised label annotations. Unsupervised deep learning turned out to be much more scalable in practice. Hence, neuroscientists at organizations without the means to train LLMs from scratch can still benefit from state-of-the-art performance by refining already pre-trained models on target tasks of primary interest, with reduced data and compute budget requirements. LLMs can thus better identify deep hidden patterns, relationships, and context within text. This led to the capability of responding to human queries, generating creative novel content, and forming accurate outcome predictions.

FOUNDATION MODELS AS COMPUTATIONAL LEGO BRICKS

Paradigmatically, LLMs are trained initially on massive text corpora, such as internet content and other public or private sources. This leads the model to develop and instantiate a general internal representation of semantic meaning, even across different languages, including syntax and grammar, although it is a matter of current debate to what extent LLMs develop an understanding of meaning.^{20,27,28} Going much beyond that, the model learns an efflorescence of general facts, certain apparent reasoning abilities, and, possibly, a semantic world representation. The evolution of foundation models can maybe be traced back to last-generation NLP models, before the transformer era (2017–), like Word2Vec⁵ and GloVe,⁶ which expressed words in continuous vector spaces (cf. [introduction](#)), hinting at universality of spanned semantic spaces.

By distilling and assimilating the quintessence from disparate expansive sources, a general-purpose representation is formed that encapsulates vast, compact, and dense human knowledge, as a form of prior for downstream modeling. It is not just memorization but information extraction and structurization. Philosophically, such a successful compression of information can

mark a milestone toward refined predictions, in that successful prediction indicates a form of information compression. Similar to a shared infrastructure or platform, such an AI engine can then act as the bedrock on which a variety of tasks can be built, making many quantitative modeling workflows feasible, efficient, and scalable. These “LEGO bricks” can be thought of as foundations because many downstream applications can be constructed on top of them, like stacking building blocks. This fresh attitude to quantitative modeling is the strict opposite of training specialized models for deployment in narrow tasks.

It is possible to use thousands of GPUs training an LLM for weeks on trillions of word tokens, with a result that can be stored and deployed on a smartphone. As a crucial consequence for the future, foundational modeling frameworks provide universal computational units that will potentially democratize access to high-quality AI solutions across broad categories of investigators. This is all the more important for the neurosciences, because investigators tend to operate on smaller datasets than those in the core machine learning community. Similarly in biological research, even the Human Cell Atlas project has produced gene expression data for “only” ~40 million human cells from ~6,000 donor individuals at the time of this writing.

Bold innovation will emerge from creative ideas on how to put these baseline operation systems to use, to revisit and tackle classic research questions—applications that were categorically unthinkable and infeasible before current NLP technology. Enabling researchers across diverse domains to bootstrap common building blocks may also help boost comparability across studies and foster collaboration across teams, institutions, and geographies. The fruit of deep learning breakthroughs will be increasingly accessible in always more resource-constrained settings. It is likely that foundation models will change the face of bioinformatics in neuroscience and biomedicine in the near future.

LARGE LANGUAGE MODELS FOR BIOLOGICAL SEQUENCES

The inductive biases of LLM learning engines immediately appear appropriate not only for word sequences, but also for different kinds of biological sequences, presenting many unexplored opportunities. The human genome encodes for ~20,000 genes, the segments of DNA that form the basis for protein synthesis in cells of the brain and other body parts. A natural proving ground, with direct relevance to the neurosciences, is the “central dogma of biology:” the one-directional flow of genetic information from (1) nucleotide sequences in DNA to (2) base sequences in messenger RNA to (3) amino acid sequences in protein products.

As a principal goal, geneticists wish to map this progression of genetic information, to link alterations in the DNA sequence itself to corresponding functional outcomes. To that end, MetaAI has presented a protein language model ([Figure 3](#)) that predicts phenotypic consequences from differences in genetic variants.²⁹ A 650-million-parameter model was used to infer the totality of ~450 million possible missense variant effects in the human genome—each a switch in a single DNA nucleotide that leads to an amino acid swap in the downstream protein

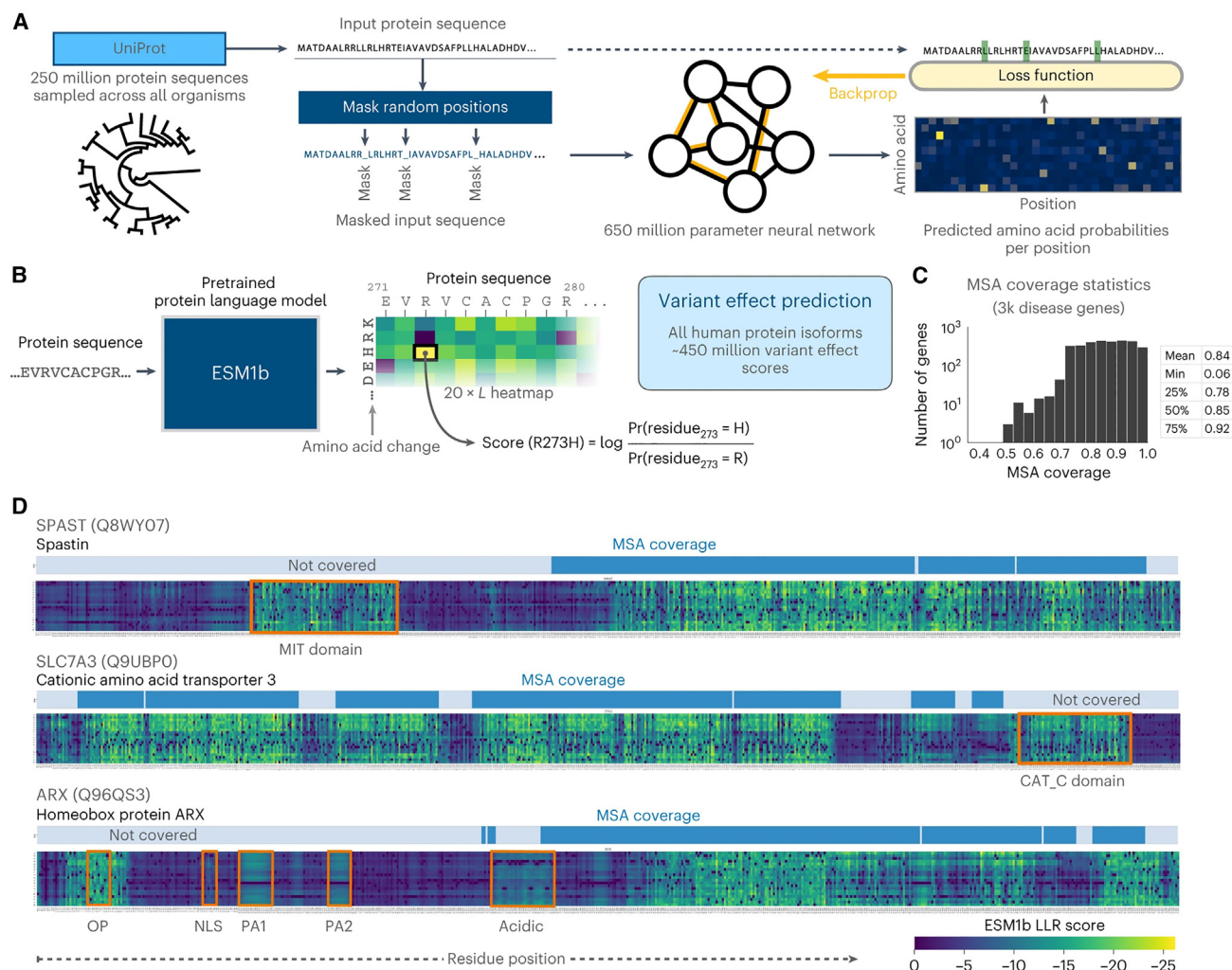


Figure 3. Protein language models to predict the functional consequences of genetic variants

A 650-million-parameter (unsupervised) deep learning model was formed from 250 million protein sequences (developed by MetaAI scientists). This framework was used to predict the totality of ~450 million potential missense variant effects (i.e., a single-nucleotide change results in the substitution of one amino acid for another in the protein produced by a gene) screening through >40,000 protein structures in the full human genome. During model training, random locations across the genome are blinded to the model and the model is trained to recover these left-out amino acids. Such modeling tools implicitly extract and represent how one-dimensional amino acid sequences lead to two-dimensional and three-dimensional features of the protein structure and function, including ligand-receptor binding sites. Such protein language models are capable of providing high-quality predictions of any amino acid sequence as well as different kinds of coding variants. Reproduced with permission from Brandes et al.²⁹

(pathogenic or benign). Such variants in DNA gene encoding are of special interest since they entail protein alterations that can be linked to disease mechanisms and possible therapeutic targets. Such approaches enable an exhaustive profiling of protein-disrupting damaging variants across the entire genome in humans and other organisms.

Can we automatically derive insights into the underlying cellular states and active biological pathways from RNA transcript expression data alone? At the level of single-cell RNA gene expressions, an LLM was trained on 10 million cells (Figure 4), each cell containing expression values for a fraction of the approximately 20,000 human genes.³⁰ As a seminal example of a foundation model (cf. above), gene sets are modeled as making up biologically meaningful processes in biology analogous to how word sets

make up meaningful sentences in language. By means of ingesting a mass of gene expression patterns, the model formed an internal representation of general principles of gene-gene relations and gene-cell relations. In addition to gene-specific tokens, special tokens were introduced to denote meta-information such as cell type, data batch, and experimental conditions like perturbations of signaling pathways and techniques used for RNA transcript sequencing. The authors also abolish the need for the input being a sequence: they designed a mission-tailored attention mechanism to get a tight grip on cohesive co-occurrence regimes of expressed genes, akin to auto-regressive generation, based on iteratively predicting expression of new genes in sets, akin to the next-word prediction goal in sentence sequences. That is, the authors recast the inductive bias from the

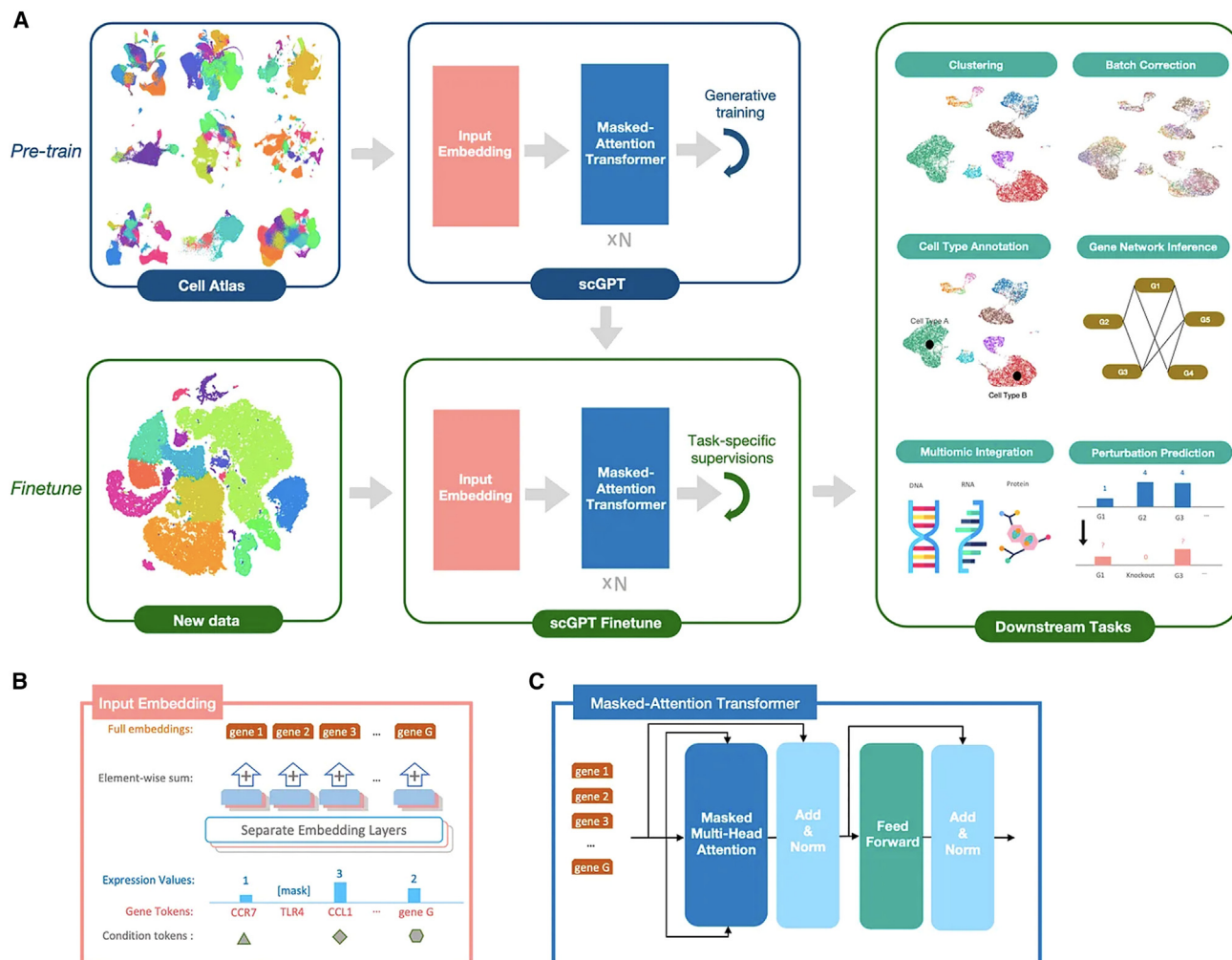


Figure 4. Creating a foundation model of the “grammar” of transcriptome biology from exponentially growing single-cell genomics data
(A) Two-step workflow of creating a foundation model for single-cell genomics: training an LLM on >10,000,000 transcriptome sequencing observations (CellXGene portal) in a self-supervised fashion (i.e., without predicting target phenotypes or classifications), the thus pre-trained LLM can then be leveraged to great effect via parameter adaptation pipelines (fine-tuning) for increased performance in specific application tasks on smaller, unseen snRNA-seq datasets (few-shot learning).
(B) Input embeddings have three dedicated information layers for gene tokens, expression values, and condition tokens for meta-information (sequencing contexts, data acquisition batches, etc.).
(C) Schematic of a mission-tailored attention mechanism for this neuroscience application. This seminal deep learning framework showcases the possibility to decipher an “underlying logic” of cell-type heterogeneity representations with their gene interaction programs and states that extrapolate to new data and research goals. Reproduced with permission from Cui et al.³⁰

usual sequence-of-words-in-a-sentence logic in exchange for a bag-of-genes-in-a-cell logic to avoid a strict sequence requirement. Once established, the trained foundational LLM could then be fine tuned and deployed with performance gains in a variety of different downstream tasks, including nuisance batch correction, cell type annotation, and prediction of targeted perturbation conditions. Such approaches show potential for leveraging self-supervised learning techniques to grasp complex single-cell mechanisms and use the ensuing internal embedding representations for integration across different organs and species.

Going from gene level to the level of 3D protein structure requires prediction of ultimate 3D configuration from 1D amino acid sequences alone. The “protein folding problem” revolves

around how information in our DNA compresses information about final protein forms. With >200 million protein structures in the database, AlphaFold2³¹ is based on LLMs to capture protein sequence interactions between amino acid residues that are far away from each other along the protein backbone. In a brute-force shotgun learning approach, the authors showed that 1D sequence information does contain key information necessary to understand the complex process of how proteins actually fold in nature. At the protein-to-function level,³² investigators trained 700-million-parameter 34-layer transformer models on 86 billion amino acids across 250 million protein sequences (UniParc database). The model-internal embedding representations were gleaned from just the sequence

information itself. The trained model was found to instantiate knowledge relevant to the ultimate protein's biochemical properties, elements of morphological structure *in vivo*, contact sites, and biological activity.

Taken together, capturing long-range interactions (i.e., tokens far apart from each other in the input sequence) turns out to be valuable to derive meaningful general principles not only in word sequences but also in different biological sequences. Nature appears to harbor underlying general rules that can now be exploited for extrapolating beyond actual sequence elements (e.g., nucleic acid, gene expression, amino acids) in the service of next-generation computational biology. The learned sequence embeddings can be used for various downstream research goals, including quality control procedures, grouping of biological entities, and enhancing phenotype predictions.

Moreover, LLMs serve as a platform that now enables advanced *in silico* models of the central dogma of biology (going from the DNA double helix to gene transcript expression to fully formed proteins). That is, once the LLM can accurately approximate the target system, reiterating trusted observations from previous rigorous experiments, investigators will be able to interrogate the LLM to extract new molecular insights about the target system and to identify broader driving biological mechanisms. We warrant caution to draw overly strict parallels between semantic language systems and molecular biology systems, given notable differences. Nevertheless, in the future, LLMs are in a unique position to help generate new sequences, never observed in the wild, that are biologically active.

LARGE LANGUAGE MODELS FOR AUTOMATED ANNOTATION

Neuroscience research often relies on accurate annotation for data elaboration, designing experiments, or interpretation of results. A recent study using classical NLP³³ explored links between brain response signals of subjects watching the movie *Forrest Gump* and the evolution of the movie story, that is, the constituent semantic facets that make up the film narrative (Figure 5). This study serves as a prime example of research that depends on pertinent high-quality annotations of data points. Regarding the brain recordings collected from the studyforrest database (<https://www.studyforrest.org/data.html>), 3,000 individual images of whole-brain neural activity were acquired from each of 15 subjects as they watched the story in the 2-h film unfold (25 such images per minute). To enrich the provided dataset, all of the scenes throughout the movie were enriched by computationally derived meta-information. To this end, text data were obtained from previously underexploited sources: time-locked subtitles and an auditory-only narrated version of the film oriented toward blind audiences that describes the events and scenes in the movie—the starting point for NLP-enabled data augmentation.

The combined scene-by-scene text information of *Forrest Gump* was captured as a bag-of-words matrix—the set of all unique words with their frequencies in a given time slice accompanying the entire movie. Latent semantic analysis was then used to deconvolve the scene-wise word statistics into unique semantic dimensions, to capture the underlying meaning and recurring themes in the story line. In parallel, in a classical top-

down approach, human annotators (a group of students) manually attached tags to scenes by choosing among a set of 52 predefined “indicators” from the audiovisual version of the movie. These choices were based on the scenes’ emotional content, circumstances, and other aspects, predefined *a priori* to be relevant, based on existing knowledge. This classical approach, which emphasized (for example) the detailed characterization of human emotions based on natural subjective experiences of human observers, turned out to miss important nuances that the text-derived semantic meaning representations reflected well, showing the potential for future LLM approaches in naturalistic neuroscience.

Going beyond the status quo of manual annotation, the NLP approach (latent semantic analysis) enabled decomposition of the story into 200 semantic movie contexts, each with their scene-by-scene relevance. As a complement to human-derived emotion annotations, the semantic contexts provided a means to track the occurrence of characters (e.g., Lieutenant Dan), contexts (e.g., war), and scene properties (e.g., day versus night). Analysis of the integrated data revealed empirical connections between brain states and specific elements, concepts, and themes within scenes.³³ Hence, algorithmically derived semantic facets were more successful in combined movie-brain-text analysis than traditional approaches that rely on human *a priori* intuition to determine which aspects should be most important.

LLMs present an opportunity to carry over knowledge and concepts from other areas of human activity into the process of how scientific research is conducted today. Pipelining annotation generation could dramatically enhance our capabilities to scale complex manual protocols such as those employed in image and video data, as in the study detailed above, but also many other forms of stimulus material, such as electronic health records, voice recordings, or the biometric outputs captured by wearable devices. Various other kinds of neuroscience-related data sources could be directly combined with brain signals on a single subject basis or at the group level. Historically, annotation of these data forms has required input from human experts, either directly or indirectly through tools dedicated to specific end-to-end input-output learning, such as neural networks trained to discern human emotions directly from visual data, or an electronic olfaction device engineered to estimate the subjective appeal of scent compounds based on their physical characteristics.³⁴ There are certain caveats associated with manual annotation in general, several of which LLMs can mitigate, including (1) high logistic and financial cost of manual effort, (2) ontological limitations of categorization systems used to derive annotation tags, (3) subjectivity from human annotators and subjectivity-based data, and (4) reproducibility.

Ultimately, as indicated above, due to the high cost required for their procurement, manually annotated vision and language datasets are relatively rare and often small in size (10,000–100,000 data points). In response to previous annotation data scarcity, numerous works^{35–37} automatically scrape readily available paired vision-text data from the internet and other general-purpose sources. Now similar feats to those achieved in the image-text annotation domain can be achieved in text-text annotation scenarios. With LLMs annotations can be automatically generated after model pre-training on a variety of data

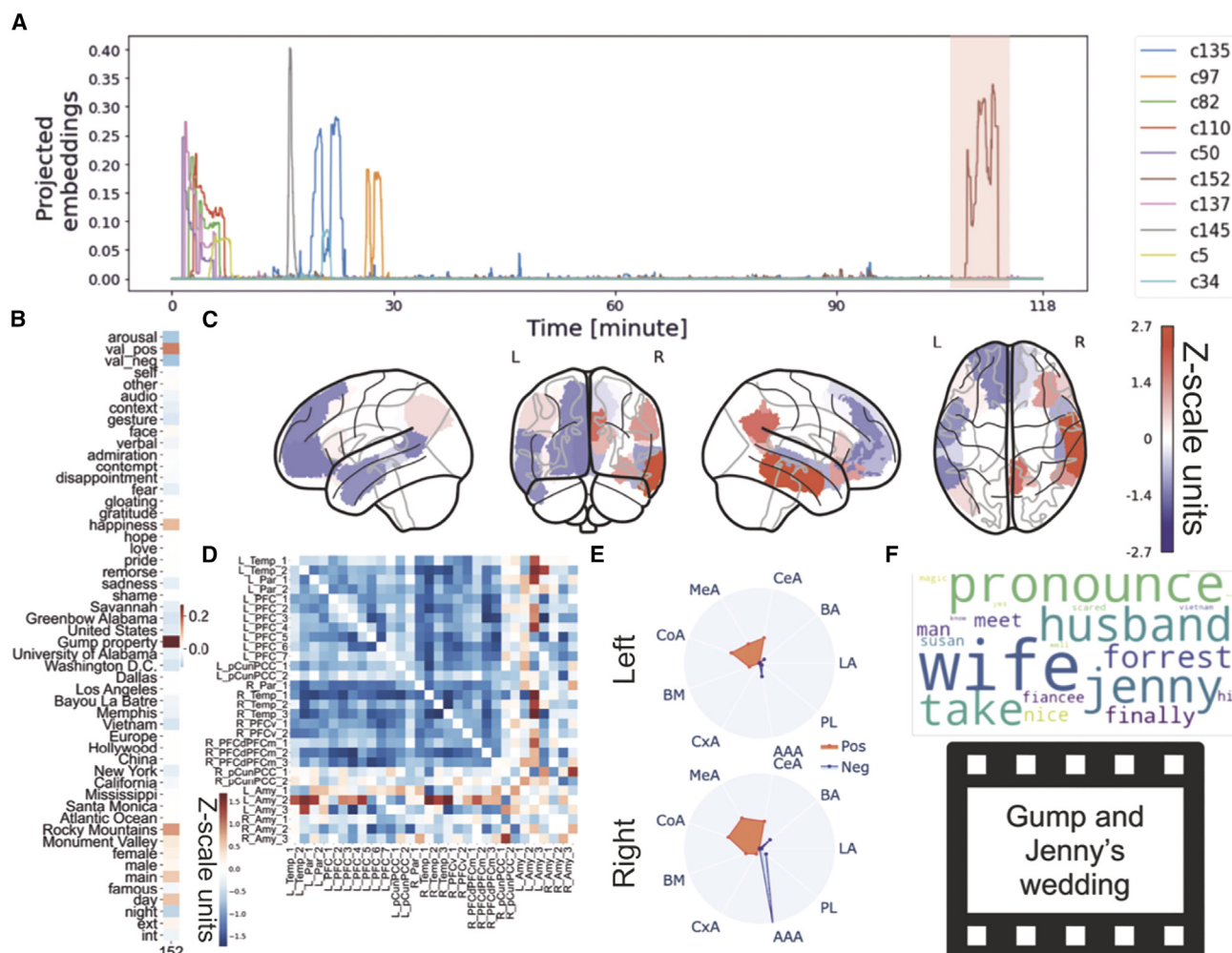


Figure 5. Multi-modal brain-text integration using NLP based on movies

Functional-brain-imaging responses were continuously acquired from 15 subjects as they watched the full 2 hours of *Forrest Gump* in the scanner. Time-locked subtitles and blind-version narration of the movie were combined in text form into word token occurrence sequences across time points.

(A) NLP algorithms allowed extraction of 200 distinct semantic dimensions that quantitatively tracked story aspects, such as characters (e.g., Jenny), location (e.g., inside a building vs. not), environment (e.g., day vs. night), or contexts (e.g., war) across the full movie material (x axis shows minutes).

(B) These bottom-up automatically derived semantic facets of the narrative were compared against traditional approach based on hand-made annotations of the movie, obtained from human raters.

(C–E) Events tracked by the 200 semantic dimensions (C) could then be analyzed for their coherent relation with neural activity changes, including their (D) detailed cortical-subcortical functional covariation patterns at the subregion level, such as (E) the amygdala. Such multi-model brain-text analyses allowed identification of which sets of brain regions are consistently linked to which sets of facets of meaning, as the narrative unfolds. Reproduced with permission from Yang et al.³³

relevant to the annotation task at hand. As a hypothetical example, a biotechnology company is interested in tagging first-hand accounts of psychoactive drug experiences with labels indicating different subjective effects; pairs of such accounts and manually applied subjective effect tags can be used for fine-tuning of a foundation model employed by the company. Alternatively, LLMs such as GPT4 can be prompted to perform this task without any additional training data, based on the assumption that their training sets provide enough context to discriminate between different subjective effect terms and examples thereof.

Phrases and sentences, like single words, can be automatically assigned information-rich semantic embeddings, the

same is true for automatically (or manually) obtained annotations. Conversion of freeform textual information via LLM “encoders” to the structured embedding vectors enables continuous quantification of discrete semantic elements. In a complementary manner, LLM “decoders” serve to transform embeddings back into language text. Preprocessing natural language as embeddings unlocks the door to new methodologies for probing correlations between distinct linguistic patterns and neural activities. Associating natural language data with neurological measurements is a step toward profound comprehension of the generation, perception, processing, and interpretation of language by the human brain. The quantitative representation of natural language text is the industry standard intermediate

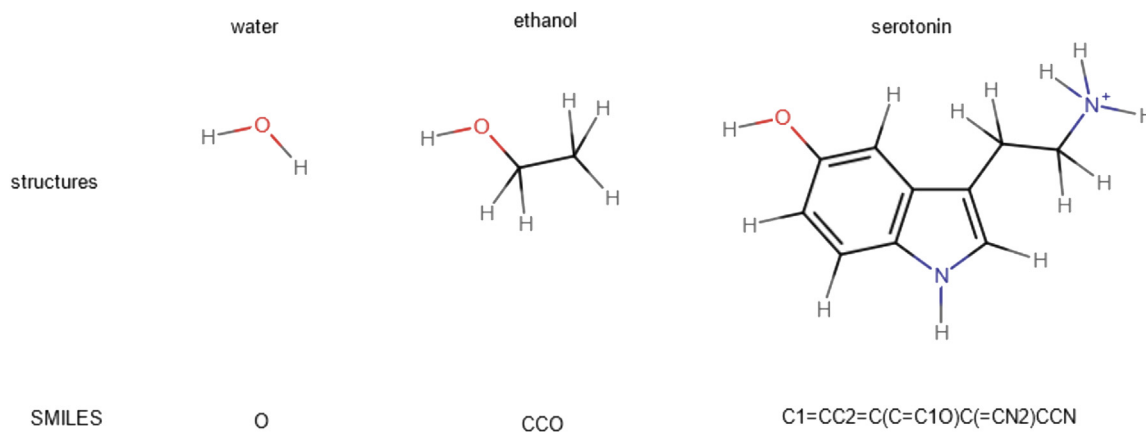


Figure 6. Examples of paired chemical structures and SMILES sequences

As a litmus test for the current state of SMILES literacy in LLMs, we prompted OpenAI's ChatGPT-4 with the following request: "For the following list of SMILES—'C,' 'CCO,' and 'C1=CC2=C(C(=C1O)C(=CN2)CCN'—please give a short summary statement about the molecule each sequence translates to." The resulting statement was as follows: "The SMILES notation is a way to represent a chemical structure in a linear text form. Here's a brief description of the molecules represented by the given SMILES strings: (1) 'C': This represents the simplest organic molecule, methane (CH₄). Methane is a colorless, odorless gas that is the simplest alkane and the main component of natural gas. (2) 'CCO': This SMILES string represents ethanol, a two-carbon alcohol with the molecular formula C₂H₅OH. Ethanol is commonly found in alcoholic beverages and is also used as a solvent and in the manufacture of various chemicals. (3) 'C1=CC2=C(C(=C1O)C(=CN2)CCN': This is a more complex molecule known as chlorpheniramine, which is an antihistamine used to treat allergy symptoms. It has a tricyclic structure with a chlorine atom, a secondary amine, and an ether functional group." This sample result indicates that general purpose LLMs like GPT-4 may have a functional understanding of the SMILES concept and accuracy in interpretation of simpler SMILES strings, like those corresponding to methane (C) and ethanol (CCO). The misinterpretation of the SMILES string for serotonin—C1=CC2=C(C(=C1O)C(=CN2)CCN—highlights the potential for future refinement of LLMs fine-tuned toward neurocheminformatic applications.

form used in computational analysis, implying reproducibility and potential for more tunable and scalable augmentation. Language, serving as a tool to encapsulate information derived from the five human senses, affords the quantified representation of a diverse range of phenomena within human experience.

Once again touching on the world of image auto-annotation as a source of inspiration for text-annotation tasks, the tool RETfound is an innovative approach that addresses the image-to-text problem in the medical domain.³⁸ RETfound is a foundation model for labeling widely available retinal images with disease categories. It is designed to expedite diagnosis of diseases including cataract, central serous retinopathy, diabetic retinopathy, glaucoma, heart failure, macular dysfunction, myocardial infarction, Parkinson's disease, stroke, and macular degeneration. The model architecture is based on the large vision transformer framework: an encoder is used to generate a high-resolution embedding space that can be used to differentiate between retinal image features, analogous to embeddings used by LLMs as a way of encoding semantics in natural language text.

RETfound's decoder is used for image reconstruction, while the encoder is used to derive features for fine-tuning toward downstream disease prediction tasks. RETfound was pre-trained on 1.6M unlabelled retinal images via self-supervised learning—a paradigm where AI models learn to find patterns within a dataset without any additional training information. For example, if a neural network was trained in a self-supervised learning task using a training set consisting of pet images, the model would most likely learn to recognize shapes that correspond to cats, dogs, and other popular pets. The model knows how to distinguish between images of different types of pets, but it does not "know" that we call one group "cats" and other information that might be linked to the pets in the images. The

same is true for RETfound in its pre-fine-tuned state: it can distinguish between distinct variations seen in retinal scan images, and this ability allows it to then be fine-tuned for particular disease detection tasks.

This fine-tuning was performed with specific expert-provided labels from a number of different datasets ranging in size. For example, the "OCTID" dataset, containing 470 retinal scans used to label conditions such as "normal," "macular degeneration," and "diabetic retinopathy," and the Moorfields Eye Hospital-AlzEye dataset, which contains ophthalmic data linked to 353,157 patients' health records who attended this hospital between 2008 and 2018, were used for fine-tuning to orient RETfound toward Wet-AMD prognosis.³⁹ With such comprehensive training, RETfound can be used to create text descriptions of retinal images based on predictions made in the context of pixel patterns in images from records generated by medical professionals. Hence, models like RETfound are designed to alleviate the annotation workload of experts, serving as inspiration for conceptual frameworks employing LLMs for similar purposes.

Image formats can be used to capture the physical world, on the one hand, and activity of neurons in the brain, on the other hand. Alternatively, they can serve as experimental variables such as pictures used in experiments depending on visual stimuli to explore links between brain scans and such stimuli. In contrast to image formats, chemical structures and descriptors thereof can capture key aspects of brain chemistry, neurophysiology, neuropharmacology, and chemosensory stimuli. Simplified molecular-input line-entry system (SMILES) is a method of description for representing and semantically re-expressing chemical structures as text-based objects (Figure 6). SMILES was first conceived based on the principles of molecular graph theory

to represent chemical structure with rigorous specification in a way that was well suited for machine processing.⁴⁰

There are vast bodies of scientific literature containing chemical names, sometimes appearing in standardized form but oftentimes not. With the proper collection, curation, and integration strategy, a corpus combining chemical names and SMILES structures could be prepared to train an LLM or fine-tune a foundation model for the purpose of exploring potential predictive relationships between chemical structure and semantic content. If this can be achieved, going even one step further, the common embedding space could be connected to a generative model that outputs chemical structures on the basis of text inputs (e.g., “I would like to see novel chemical structures that will be able to enter the human CNS, please!”). In a not-so-distant future, such multi-modal LLMs could become a valuable partner to scientists to enhance the creative process of generating entirely new molecules with targeted properties, whether they be physical, chemosensory, or pharmacological.

Another potential use of a common embedding space between SMILES and natural language would be to analyze mixtures of chemicals as opposed to single chemicals. Just as interpretation of words and phrases appearing in natural language can be significantly influenced by their context, perception of odorant molecules present in chemosensory stimuli (naturally encountered as mixtures) is influenced by the combinations and concentrations of other mixture constituents. Furthermore, small molecules such as neurotransmitters, hormones, drugs, and toxins often act in tandem with their metabolites, impurities, and other biomolecules. These combined elements can exert biochemical and physiological effects in their surroundings, such as binding to target receptors or modulating signal transduction pathway activity. Hypothetically, the common latent embedding space of an LLM that was trained with SMILES and natural language could be used to navigate the complex, context-dependent, multiplicity of action by chemicals and mixtures, of direct relevance to neuroscience.

Another issue presented by annotation, separate from high cost, is the fact that annotation that relies on a predetermined ontology or classification system will be limited by the descriptive capability of such a description system. Typically, individuals who perform annotation tasks must be trained to properly use a given ontology for application of classifications to data points, as an attempt to mitigate the known challenges of inter-rater variability. Sometimes the training required to properly annotate data is extensive, and the annotators must be qualified as subject matter experts as opposed to lay persons. The embeddings generated by LLM encoders can potentially be “translated” to a set of terms within a targeted ontology using techniques like semantic similarity measurement or clustering.

If left untranslated, LLM embeddings offer a high level of semantic granularity that is not afforded by ontological classification. This specificity is valuable in any instance where researchers are interested in recording distinct outcomes, because it enables flexibility in terms of how to categorize annotations in a way that is of direct relevance to the specific experiment to be performed. As a straightforward hypothetical example, one could (1) generate semantic embeddings from annotation labels or another experimental variable recorded via

text, (2) generate embeddings from terms present in targeted ontologies, or (3) calculate cosine distance between the two sets of embeddings to identify the “nearest neighbor” term from targeted ontologies for each text-based experimental variable. While such an approach might not afford the accuracy that can be provided by subject matter experts, what it lacks in resolution it compensates for with objectiveness and operational consistency that increases both scalability and reproducibility of annotation at scale. On the other hand, the embeddings yielded by LLMs also provide researchers with a means to analyze annotated datasets, via clustering or more sophisticated techniques, to identify new classification systems.

Ideally, we will soon be able to assign expert-grade annotations by means of LLMs, even in the absence of close collaboration with a card-carrying domain expert. To make matters more interesting, once it has been demonstrated that LLMs can apply pre-existing ontologies for annotation in a manner that is comparable or superior to the performance of experts, we can turn to “expert LLMs” to help with the identification and validation of new terms and ontologies that are derived in a data-driven way. We can also examine the results of LLM-based annotation to challenge incumbent classification systems that were designed by limited heuristics. While rule-based solutions operate based on explicit predefined criteria, black-box AI solutions (despite their opaque decision-making processes) often excel in handling vast and complex datasets,^{41,42} achieving superior predictive accuracy where traditional methods might struggle. Using LLM-assisted annotations as a complementary approach to supplement legacy top-down (e.g., manual categorization by domain specialists) and standard rule-based (e.g., predefined algorithms for data point classification) solutions is one way we can simultaneously leverage the knowledge that comes from expert experience as well as the new insights we can obtain from LLMs, a data form that can truly “speak for themselves.”

LLMs have been described as chameleons (<https://karpathy.ai/lexicap/0215-large.html>) or enabling forms of “role play.”⁴³ They can take on the personality and adopt the thought or writing style of known persons or categories of persons with specific traits, such as Charlotte Brontë, Carl Sagan, or a neuroscientist. This capacity can be leveraged in a variety of ways. In some annotation tasks, it is beneficial to seek counsel with a panel of experts across multiple disciplines, as opposed to a panel of evaluators who all share the same background. Several LLMs assuming different “chameleon” stances could be used in parallel in an annotation task, analogous to a panel of human raters. The LLMs can be asked to take the position of different experts, personality types, professions, age groups, and cultural backgrounds. LLMs not only address the problems presented by the influence of individual subjectivity on annotation tasks, but they also simultaneously enable expression and manipulation of such subjectivity. LLMs can eliminate the transient fluctuations of ephemeral emotional states experienced by human annotators, and if called for they can introduce them in a controlled and repeatable way.

There are many sources of inconsistency in the language used to describe neuroscience research as well as subjective experience. These vagaries fuel disagreement in the interpretation of annotation between different researchers. The universal

format of coherent semantic embedding spaces enables capture and manipulation of vague or subjective language. Crucially, these representations are exactly repeatable across laboratories and other contexts of research and analysis; so long as the same LLM is used for the same task, with the same totality of previously estimated model parameters. From a practical standpoint in regards to scientific research, this feature should have a significant effect on shareability of annotated data across investigators or labs, hopefully expanding the breadth and depth of downstream applications for auto-annotation of datasets via LLMs.

Separate individuals can annotate the same data differently, and even a single annotator's responses may vary over time. LLMs offer a more stable and consistent form of annotation because they are trained on a broad dataset and are not influenced by subjective experiences. In place of the subjectivity that influences human performance on manual annotation tasks, LLMs have a nuanced mapping of linguistic context captured by the use of language in their training corpora. Off the shelf, LLMs can be thought of as approximations of the average mind of all internet users, "crowd-sourcing thought," since a large portion of their training corpora is derived from the internet. Alternatively, in the event that a foundation model does not appear to capture enough nuance to achieve a specific task, it could be fine-tuned to approximate the average mind based on a certain subset of websites or internet users.

There are typically subjective aspects to the process of manual annotation, especially in the case that the object being annotated is something experienced on a subjective basis. In the example above where students were manually annotating scenes from *Forrest Gump*, they were asked to annotate the emotions they perceived to be expressed by the actors in the film. This task requires subjective interpretation of emotions portrayed in the movie in the first place, on top of the fact that emotional experience is highly subjective in nature. The study-forrest dataset also includes annotations of the physical location in which each scene takes place.³³ Even though these annotations ("night" vs. "day," "inside" vs. "outside") are largely objective judgments made by subject matter experts (two individuals with academic background in film), there is still room for subjective interpretation by the annotators, as exemplified by the operational definition of "day" as any scene that was illuminated by sunlight, as opposed to some other determining factor.

LLMs enable reconciliation between the world of subjective phenomena and objective measurement. The representation of semantic entities via LLM embeddings preserves the discrete subjective or contextual meaning in text such that it can be compared in a consistent way with other text. For example, imagine a scenario where sentences are collected from social media posts to be auto-annotated with labels indicating emotion to be used in a training set for an NLP model that predicts the emotion of social media users from their posts. No matter how unique each envisioned sentence is, the distance between their embeddings and the embeddings of terms such as "enthusiastic," "depressed," "nostalgic," or "peaceful" can be calculated in a uniform fashion. Due to the fact that LLM training corpora capture a large volume of text describing subjective phenomena, more stable and consistent annotations yielded by LLMs can

readily be used to characterize subjective experience-based data elements, without using subjective human judgment as part of the annotation process.

The use of LLMs to automate annotation tasks is not a step-wise improvement; it is a next-generation approach that can disrupt a mainstream practice that would otherwise continue to fall prey to subjectivity and other forms of idiosyncrasy. For instance, consider the task of annotating emotions in a collection of diary entries. If given to a group of human annotators, one might label a passage as "sad" based on their personal experiences and cultural background, while another might see it as "reflective" or "nostalgic." Because LLMs are autoregressive, state dependent, and have hyper-parameters such as temperature (cf. previous section "[data science perspective on large language model solutions](#)"), there are reasons why the answers for an identical prompt may not necessarily be *exactly* the same. Nonetheless, if experimental conditions are held constant, answers from LLMs should be largely restricted to narrow regions of semantic space. In this way, the LLM may offer a level of objectivity and consistency that human annotators, with their inherent subjectivity and idiosyncrasies, simply cannot match.

LARGE LANGUAGE MODEL FOR TEXT SUMMARIZATION AND KNOWLEDGE INTEGRATION

The wide-ranging ballpark that is neuroscience touches various disciplines from physics to psychology. This wildly interdisciplinary field produces a myriad of rather separate experimental findings that can be overwhelming to integrate by human effort alone. Moreover, the breadth of the field often results in researchers working within a particular sub-community, focusing on narrowly specialized research areas, and potentially missing out on opportunities through potential cross-fertilization with other sub-disciplines. There may also be certain tasks that go beyond human cognitive ability, including reading experimental results that contain immense numbers of datapoints or distilling the content of all major scientific publications from the past year. LLMs can assist researchers to absorb large bodies of text that would otherwise be challenging to read and absorb in a reasonable amount of time.

The capabilities of LLMs extend beyond typical text summarization tasks, where text being gathered is presented as human-readable (albeit lengthy) natural language. LLM embeddings provide an objective quantification of subjective text (cf. previous paragraph) to resolve linguistic ambiguities and standardize outputs. Subjectivity-based text could be simple words or phrases such as those used to capture emotions portrayed by actors in *Forrest Gump*³³ or those used to describe chemosensation of odor or flavor compounds.⁴⁴ Or they could be far more complex, as is the case for text used in psychedelic research.

The common expression "the psychedelic experience" is used in a way that implies a uniformity across "trips." In actuality, the psychedelic experience is full of nuance and variation rooted partially in the drug user's set and setting as well as partially in psychopharmacological differences between drugs. Understanding the underlying factors determining nuanced outcomes observed in psychedelic drug users should help us to understand whether certain drugs or varieties of subjective effects

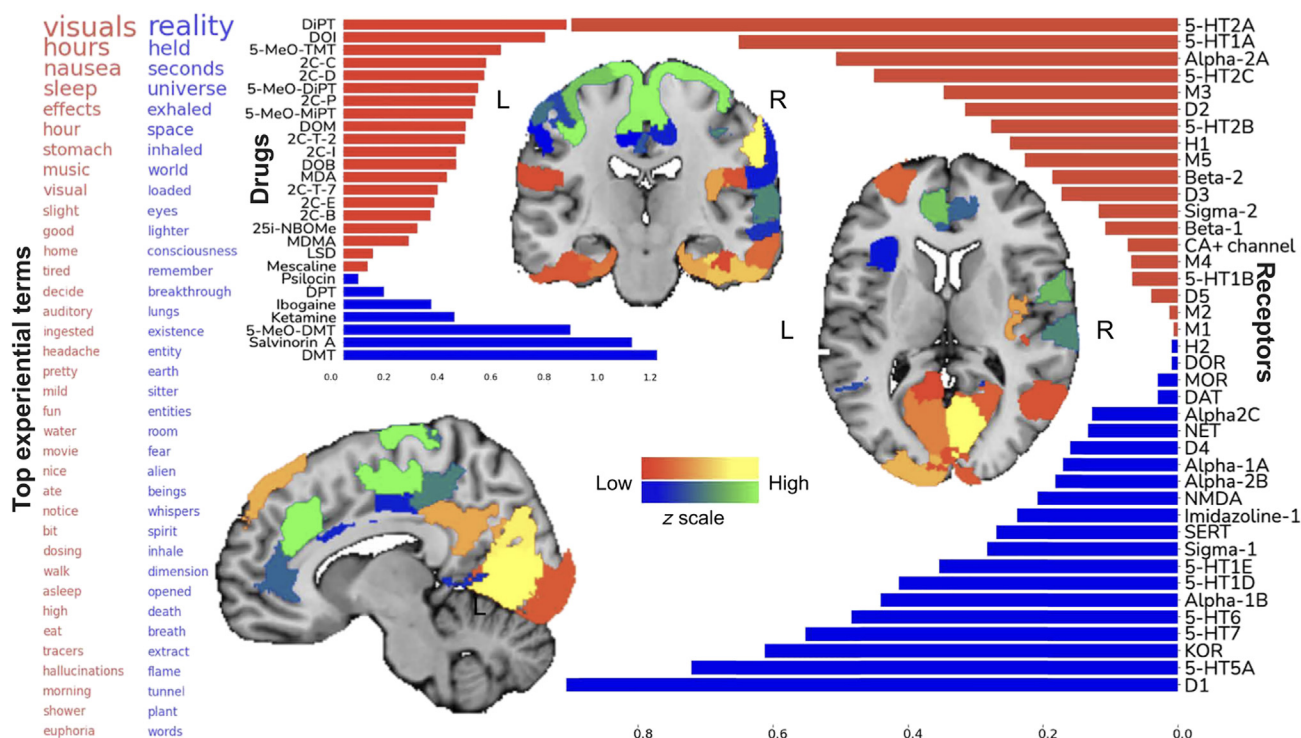


Figure 7. Multi-modal receptor-text integration using NLP to reveal the mechanistic basis of psychedelic drug experience

Mega-analysis of >6,500 text reports of hallucinogenic experiences induced by 27 different drug compounds for integration with drug-specific receptor binding profiles of 40 neurotransmitter sub-systems in the human brain. Free-form text reports from drug users were recast into word occurrence matrices and analyzed for coherent semantic themes using NLP algorithms. To achieve cross-modal synthesis, receptor-experience patterns were derived as sets (linear combinations) of neurotransmitter systems and sets (linear combinations) of usages of vocabulary words. A ranking of most relevant words was derived for each particular receptor involvement profile. This unprecedented fusion of kinds of information sheds new light on changes in conscious awareness triggered by drugs acting on dedicated transmitter receptor systems. Reproduced with permission from Ballentine et al.⁴⁵

can be employed to treat specific conditions, just as the different experiences provided via ingestion of psilocybin and MDMA have each shown early success in the treatment of OCD and PTSD, respectively. To investigate such nuances, a recent study used NLP techniques to analyze 6,850 “trip reports” from psychedelic drug users (Figure 7). The objective of the study was to draw connections between subjective experiences, 27 qualitatively distinct drugs, and a set of 40 associated neurotransmitter receptors expressed in the human brain.⁴⁵ The results of this study include detailed word lists of ranked relevance for semantic dimensions capturing major themes present in experience reports, derived via canonical correlation analysis (CCA).

Human interpretation of the complex themes captured by thousands of words in a particular order is quite difficult. Each word in the ranked lists provided via CCA carries its own potential for subjective interpretation. The varied range of potential interpretation is further widened by the context provided by neighboring terms, as well as the transition in general meaning captured by different subsections of the lists (i.e., top 1 percent versus top 5 percent). Despite the results being presented as dense lists of highlighted words, an LLM can seamlessly abstract away from these word sets by extracting semantic core themes from them, deriving shared higher-level categories of subjective effects elicited by psychedelic drugs. These higher-level categories can then be leveraged to interrogate new hypotheses

for drug discovery platforms and experimental treatment approaches, in search of new psychedelic drugs with targeted subjective effects intended to treat specific conditions. Future use of LLMs highlights yet another opportunity for researchers to glean insights from complex, unstructured data that humans might find challenging to cope with.

Medically oriented LLMs, such as Meta’s PMC-LLaMA,⁴⁶ offer a promising solution to the need for sifting through extensive text sources to aggregate and synthesize the essence of their meaning and informative value. By gathering and summarizing vast information landscapes, these models provide access to the quintessence and perhaps elements of understanding of complex topics. Specifically, PMC-LLaMA was designed to support individuals in navigating vast swaths of medical information by training on a massive corpora: 4.8M biomedical academic papers, 30K medical textbooks, as well as 202M tokens of medical question-answer pairs, rationales for decision making, and conversational dialogues. PMC-LLaMA was shown to be convincing in producing reasonable and coherent responses in zero-shot assessments of medical knowledge prompts, for example, answering questions from patients about their urinary tract infections and in-depth exam questions about microbiology and pharmacology. When asked a multiple-choice question about a drug-drug interaction involving tuberculosis and hormonal birth control medications, PMC-LLaMA correctly indicated the mechanism of

the drug-drug interaction and elaborated on the rationale used to arrive at the answer (CYP3A4 induction via antibiotic drug rifampin leads to decreased concentrations of hormonal birth control, ultimately increasing the possibility of an unintentional pregnancy). PMC-LLaMA underscores the effectiveness of data-centric approaches in specialized domains and the value of domain-specific model tuning.⁴⁶ Such impressive responses to prompt queries represent a scenario of machine-assisted human intelligence where LLMs can be tailored to effectively educate users in specialized areas, highlighting the potential society-transcending impact of such models and the importance of domain-specific model development.

As another concrete example of gains in everyday life, instead of plowing through a series of thick textbooks, a medical student preparing for an exam could query models like PMC-LLaMA for information regarding specific topics to ensure they cover a wide range of material in a way that is more time efficient. Just as automation in industry made more time available for workers to accomplish other tasks, we can expect to see similar opportunities presented by LLM-based developments. However, the improvements will not all be simply life enhancing; many applications, such as an interactive LLM that has access to patient electronic health records, could potentially be lifesaving. Unfortunately, a recent statistical investigation by Rodziewicz et al.⁴⁷ estimates that ~400,000 hospitalized American patients experience some type of preventable harm each year, with roughly a quarter of such cases resulting in death. The lifesaving potential of AI in medicine has several areas where it can shine, such as (1) reducing workload of medical professionals, so that they can be more efficient with their time to better assess and treat their patients, and (2) acting as an early warning system to alert of potential adverse events from ranges of available treatment strategies.

MULTI-SOURCE AND MULTI-MODAL LARGE LANGUAGE MODEL SYNTHESIS

Over the last decades, the neurosciences have expanded into increasingly segmented silos of research activity. For example, Alzheimer's disease (AD) is studied in several largely disconnected research communities. Epidemiologists studying the etiology of AD disease in human population strata do not regularly talk to geneticists, practicing neurologists, brain-imaging investigators, or animal experimentalists. The geneticists studying genome-wide risk variants related to AD do not necessarily cross-reference or integrate existing knowledge from these other neuroscience communities either. The imaging neuroscientists devoted to structural and functional differences in the AD brain do not necessarily take into consideration aspects of epidemiological population stratification when designing and interpreting their studies, and so on and so forth. Each AD research community operates in what appears to be its own "bubble," with its own set of notable scientists, its own pool of commonly entertained hypotheses, and its distinct process of knowledge accumulation, yielding large quantities of papers published per year.

Given increasing amounts of research output every year, a single human is increasingly unable to read all these papers. Many areas of neuroscience research activity are siloed in similar

ways. Such knowledge fragmentation is maybe one of the biggest challenges of the scientific enterprise in the 21st century. LLMs now offer an opportunity to assimilate and translate expanding knowledge from several complementary viewpoints on a single neuroscience topic.

LLMs are also starting to get tailored toward the medical domain, with promising results in tasks like medical exams and record keeping. To date, AI in medicine has often been based on computer vision tasks, with limited integration of text, voice, and other kinds of information. Summarization and integration of various data sources through LLMs thus holds tremendous promise for advancing AI assistance for practicing healthcare professionals. Biosensors, genome profiles, medical records, patient testimonials, metabolic panels, and other laboratory assays are examples of potential data sources for building a multi-modal AI framework oriented toward elucidation of patient-personalized clinical pathways.⁴⁸ The potential for such AI solutions is vast in terms of the direct impact it could make in the lives of patients and the performance of medical professionals, and has not been fully realized yet.⁴⁹

Currently, the use of LLMs in tools designed to help lighten the annotation workload of professionals is also a subject of interest in the realm of medicine. Although the ethics of using LLMs in medicine and medical research are beginning to be discussed,⁵⁰ it is now becoming apparent that LLMs could be effective as adjuncts to processes that currently occupy a large amount of human time and effort, such as electronic health record creation and processing, as well as many other activities such as diagnosis and prognosis of disease.

As a next holy grail, what non-text data modalities can be made LLM actionable? Broadly, LLMs may very well be the first technology that can seamlessly combine structured and unstructured information, both dynamically and at scale. Moreover, ChatGPT and similar LLM variants have successfully aggregated disparate text sources from several languages, geographies, and cultures into a single model instance.

LLMs hold promise in bridging the gap between disparate kinds of information, most obviously perhaps computer vision (i.e., images) and language (i.e., text). As a recent example from the machine learning community, Alayrac et al.³⁵ demonstrated that including such additional modalities can improve language modeling. Flamingo models were trained on large-scale multi-modal corpora, drawn from the internet, containing naturally contextualized text and image information. Ensuing few-shot learning capabilities can be adapted to a diversity of tasks involving image and video material. Subsequently, the models can be queried to enrich other image material by generating free-form content or checking of predefined multiple-choice questions. Prompting a Flamingo model with task-specific examples can provide practical benefits in many settings, based on visually conditioned autoregressive text generation. As an early example in the neurosciences for reading the image from the subject's mind,⁵¹ a study used a model to attempt to reconstruct natural images only from brain activity measurements.

Further, DALL-E/CLIP (made available 2021/22 by OpenAI) was an early example of next-generation text-image fusion in generative AI, initially using GPT-3 variant under the hood, aiming at always more realistic images generated from user

prompts. This multi-modal fusion engine can synthesize various forms and styles such as realistic natural images, painting-like art and symbols, and internal models of design schemes, invoking real and imagined objects, scenes, and people, without close training examples (zero-shot learning). Its component CLIP (contrastive language-image pre-training) was trained on ~400 million pairs of images and text captions from the internet. This model is used to subselect among images generated by DALL-E for optimal output generation. CLIP combines computer vision and NLP within a single network to deep process, categorize, and generate text annotations for a wide array of images. Without a strict requirement for task-specific training, it can generalize beyond its specific training information toward new, never encountered tasks (cf. transfer learning above). In the neuroscience context, several forms of “images” could potentially be ingested in future LLM frameworks, such as structural and functional MRI brain imaging, PET, and fNIRS, and more broadly also EEG/MEG-derived brain images.

Hence, an important future research avenue will explore to what extent DALL-E/CLIP and similar emerging technologies can successfully extend from natural images to different modalities of brain “images.” For example, the NeuroSynth database presents a bottom-up approach⁵² that has automatically extracted the activation coordinates in 3D image space for >3,000 published articles of brain-imaging task experiments, together with the full text of those articles. As such, this initiative has completed the effort of assembling a corpus of image-description pairs that has already provided value to the neuroscience community through a web interface for user queries. In a parallel research stream, the BrainMap database^{53,54} has devised a human-made ontology of mental categories at play in brain-imaging experiments in a top-down fashion. The description systems for cognitive phenomena were hand-designed by human domain experts. Here too, an existing effort has already aggregated image-description pairs that may serve as an attractive starting beachhead for training or refining state-of-the-art multi-modal LLMs. One idea would be to merge NeuroSynth and Brainmap based on the studies that are available in both databases, expert definitions and full text annotations complementing each other to enable LLM-empowered queries and perhaps reasoning across both kinds of brain image meta-information. More broadly, such avenues aimed at transcending content types are especially promising because LLMs offer an unprecedented opportunity to fuse structured and unstructured information in a unified framework.

Over the next few years, neuroscientists can systematically examine what kinds of brain-relevant information lend themselves to emerging modes of LLM integration. What kinds of neuroscience information can be tokenized, and how? Recent LLM research showed promise for leveraging of embedded amino acid blocks, genes, and their mRNA transcripts, cells and cell types, phenotypes, and disease states. Potentially expanding their applicability in biology and medicine, LLMs may also be able to process token-transformed instances of brain region activity, white-matter fiber pathway involvement, brain structure change location, frequency band change in EEG/MEG, or calcium imaging. In so doing, neuroscientists may bring together sequence semantics across datasets and biological

perspectives, forming unified perspectives on the brain. This agenda may demand model architecture innovations to represent these information layers. Alternatively, we can use the outputs of generously pre-trained LLMs as a form of distillation by encoding specific information modes, for integration in subsequently trained smaller models for an ultimate research goal. Concretely, datasets from the UK Biobank and other mega-datasets allow the LLM to associate genomic variant information and other molecular data with a variety of human health information. As a core aspiration for the intensely interdisciplinary neuroscience endeavor, LLMs can help us bridge the divide between disparate neuroscience communities and can enable us to form NLP models that can amalgamate the world’s knowledge mosaics.

EPISTEMOLOGICAL AVENUES TO OVERCOMING THE CURRENT CONCEPT CRISIS

LLMs may provide an alternative toolkit that turns out to be valuable to audit and edit the human-conceived notions that neuroscience investigators act on to understand the brain. It is important to appreciate that, especially in classical hypothesis-driven research, the entire research endeavor hinges on the pre-assumed validity of the cognitive and neural terms that were used to articulate the experimental research conditions. Yet, many cognitive or psychological terms in frequent usage have brittle definitions and cannot be directly observed in nature. Many human-expert-determined concepts in neuroscience may not denote “natural kinds” in that they do not carve out discrete neural circuits in nature. An overwhelming majority of concepts of cognitive processes have been coined well before neuroscience emerged as a coherent discipline (around the middle of the 20th century), before brain function was beginning to be understood. Further, certain behaviors or cognitive concepts may emerge only in carefully designed experiments in healthy subjects or clinical conditions such as patients with localized brain lesions.⁵⁵ According to this view, neurocognitive processes can be decomposed during subject engagement in specific experimental tasks, as an avenue to revealing the mapping between brain and behavior. Perhaps it is about time that we put to a test their validity with a disciplined data-driven approach.

The intricacies neuroscientists face when articulating observation of phenomena in the brain appear closely related to Ludwig Wittgenstein’s second main book *Philosophical Investigations* (1953/2001). The late Wittgenstein argued that confusion introduced by human language itself is the origin of most philosophical problems. For example, in psychology, there is still no widely accepted definition of even simple words like “cognitive” and “emotional.”^{56,57} Further, the brain network consistently recruited during theory-of-mind cognition, taking another’s individual’s point of view, is also consistently involved in an array of diverse psychological processes, including moral thinking, autobiographical memory retrieval, and spatial navigation.^{58–60} Our legacy catalog of neurocognitive frameworks may not go in the right direction.⁶¹

For example, why do we implicitly expect that terms and notions of William James’ opus magnum (*Principles of Psychology*,

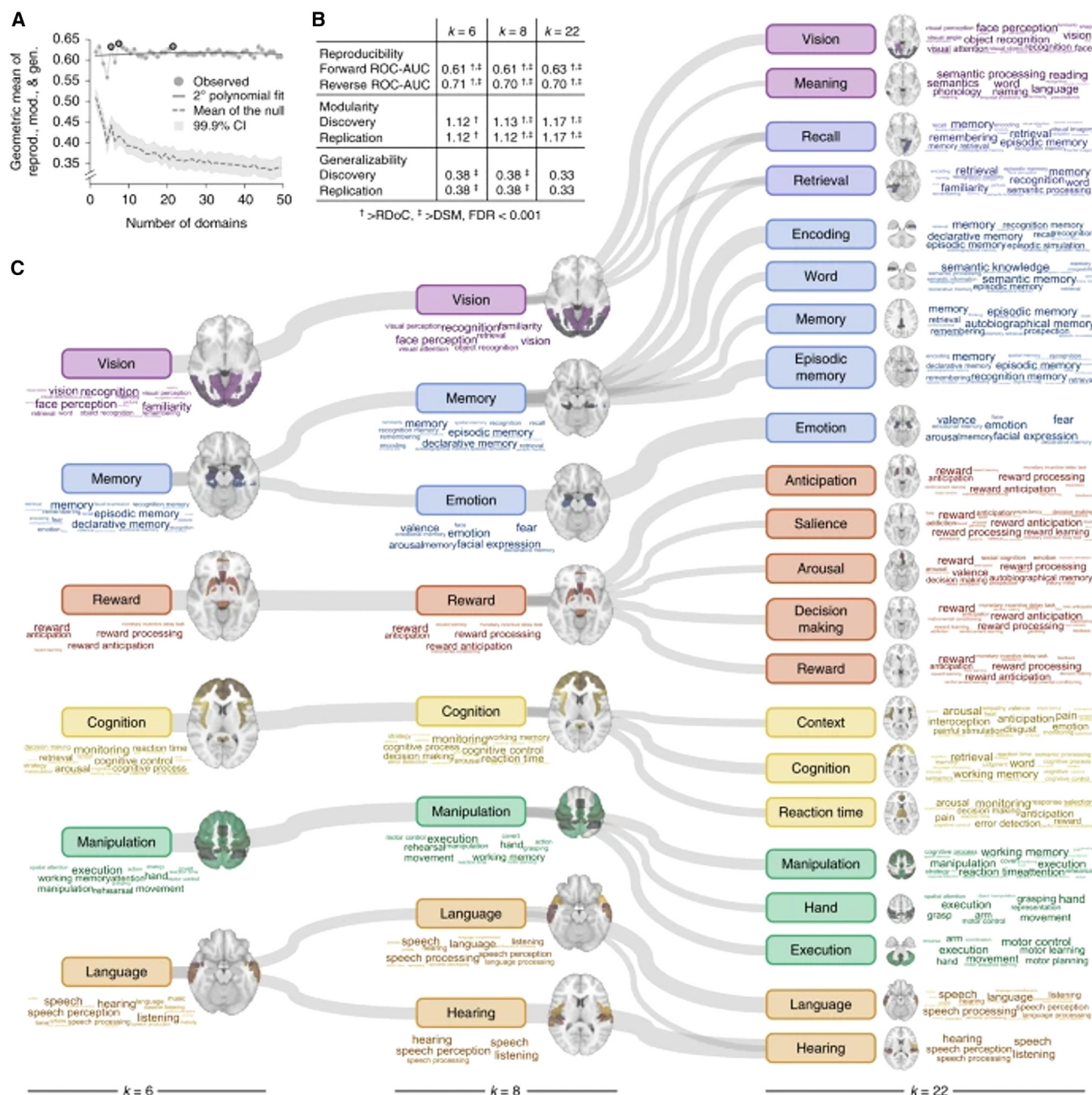


Figure 8. NLP tools to organize the existing knowledge of human cognition in a fully bottom fashion

What is the modular architecture of mental constructs? Based on several performance metrics, partitioning brain-behavior components into 6, 8, and 22 notions was validated to give the best results. This approach revealed a coherent hierarchy of candidate cognitive ontologies, pooling information from a corpus of thousands of brain scans from published stimulus-response experiments. The number of mental function terms was determined by how well term usage in paper articles predicted evoked neural activity responses (forward inference) and how well these functional brain circuit changes predicted the term usage (reverse inference). This seminal study yielded several findings that are in conflict with canonical psychology literature and widespread teaching at universities worldwide. As such, to oppose the incumbent expert-designed terms to denote cognitive processes, there is now ample opportunity for next-generation text mining tools to let the data speak for themselves, thus carving out fresh boundaries around functional domains of the human mind. Reproduced with permission from Beam et al.⁶⁷

1890) denote unique brain mechanisms? Further, when we encounter challenging-to-reconcile findings, we sometimes have the tendency to make up a new term instead of really getting at the core of the problem. Many neuroscience investiga-

tions take the outside-in approach⁶¹: they make up concepts first and then, only as a second step, go about locating or characterizing them in measures from the brain. This closely relates to what has been called “neo-phrenology” by some authors—

BOX 1. Limitations of current LLM tools

Despite LLMs being perhaps the most rapidly evolving technology of all times, a number of challenges remain for today's incarnation of these models.

Hallucinations: Refers to the common issue where the model generates text or information that is not anchored in reality or the provided context. The model might generate plausible-sounding but incorrect or fabricated information, despite a deceptively confident tone. By design, an LLM generates text, whether or not the model is certain or uncertain of its output. Hence, current LLM variants may be inherently less well positioned for accurate and reliable information queries (e.g., give exact paper references).⁷⁰

Dependency on big data: LLMs have a hunger for vast quantities of input data. Extended fractions of the internet have now been exploited for LLM development. Consequently, one might wonder whether we have saturated our available training data already. What are modes of future data generation for training ever-more powerful LLMs? One possibility is that last-generation LLMs will increasingly generate output data, on the internet or other venues, which will be fed back into next-generation LLMs. It is currently hard to anticipate what the ramifications of such a recursion scenario would be. As one possible consequence, solutions to performance benchmarks may increasingly contaminate the training data.

Resource hunger: Deploying LLMs requires a significant amount of computational power, information storage capacity, and energy consumption; probably also in terms of the lasting environmental footprint.⁷¹ For the goal of training LLMs from scratch, the extent of necessary richness in compute-storage resources probably sidelines the large majority of institutions in industry, academia, and government on the planet.

Reasoning: Instances of this model class often lack common sense or the ability to understand and respond to novel situations that were not present in their training data. How do we make sure that LLMs act in line with human values (the so-called alignment problem)? Also, at times, these models may generate text that is not relevant or not fully aligned with the context provided to it. As part of an explanation, LLMs perform fairly well in single-step reasoning tasks, but face challenges in sequential integration of consecutive reasoning steps.

Biases and other ethical considerations: Further, LLMs inherit the biases that may be present in the ingested datasets during training. The models can inadvertently generate harmful, offensive, or otherwise skewed outputs.⁷² Reinforcement learning from human feedback, calibrating LLMs toward the kinds of answers that humans expect, may be part of the solution. Further, current LLMs do not necessarily work well across languages and cultures.⁷³

Certification: Watermark assignments or deciding whether a text was LLM generated or not is probably challenging to impossible.

Lack of explainability: For users and developers alike, it remains difficult to understand why a given model generated a particular response, which is a significant limitation for applications that require interpretability and transparency, especially given increasing political pressure for white-box machine learning solutions (cf. GDPR law in the European Union). Closed-source LLMs further complicate this matter.

Diminishing returns in scaling: Continuing to increase the data quantity and compute/storage resources is already starting to hit regimes with diminishing returns. Alternative strategies for bringing the emergent abilities of LLMs to the next level will probably be required in the future.

the reductionist approach or “overlocalization,” attempting to map terms onto local geographies of the brain.⁶² While modern neuroimaging has shown that specific brain areas are indeed more active during certain tasks, the brain is highly interconnected, and many cognitive functions are distributed across networks. Thus, pinpointing a single “spot” for a complex function can be misleading.

The research focus should perhaps be placed on the actual responses of the brain, not the (reification of) human-invented terms themselves. Indeed, it is neurocognitive processes in the brain that give rise to behavior and cognition. In short, it remains elusive how and to what extent psychological terms map onto regional brain responses, and vice versa.^{62–64} For these reasons, some authors put forward that neuroscience is increasingly rich in data but remains poor in theory,⁶⁵ pointing to the acute need for new means of generating research hypotheses.

The analogous point can be made about definitions of brain disease, and especially terms in psychiatry. A same notion is not uniquely related to the same mechanism, and the same mechanism does not often isolate a distinct diagnostic entity.

This realization may be part of the reason why an identical drug class often helps alleviate symptoms for nominally separate psychiatric conditions. The DSM-5 and ICD-10 manuals catalog psychiatric diseases based on the judgment of selected experts. Moreover, funding agencies entrust scientists with funding commitments only if their research proposals' rationale and expected outcomes are firmly rooted in these human-made diagnostic categories. It has, however, become increasingly clear that pathophysiological processes in primary biology are exceedingly heterogeneous and mutually overlapping in degrees, even at the raw genetic level.⁶⁶ Consequently, today's description systems for mental health conditions help communication between practicing medical doctors but lack biological validity in research and predictability in clinical care.

Despite these evident shortcomings of incumbent description systems in neuroscience, there have been few attempts to build such a system of semantic notions in a bottom-up fashion. In a seminal study (Figure 8),⁶⁷ a data-led approach was devised to design a framework for neurocognitive categories by pooling across information from ~20,000 brain-imaging papers in

humans. Capitalizing on the accumulated data trove of >25 years of brain-imaging research, NLP algorithms mined the semantic content of the research articles, which was interfaced with >600,000 topographic locations from functional brain scans (fMRI, PET). Paying equal and simultaneous attention to both semantic principles and neural activity principles allowed a systematic integration of brain and behavior in a holistic approach. Among other benefits, this approach helps overcome the dilemma between the forward-inference (concept-to-brain reasoning) vs. reverse-inference (brain-to-concept reasoning) that is haunting the neuroscience enterprise.⁶² In empirical validation analyses, such “computational ontology” was demonstrated to better reproduce the term-function links in new, unseen research articles than widely embraced description systems in neuroscience and psychiatry.

Taken together, the narratives and stories that we use to describe the world shape the way we design our neuroscience experiments and interpret what we find. In neuroscience, true progress requires particular sensibility to word usage, language hygiene, and variants in conceptualization. In a future of LLM-empowered neuroscience, we may be able to reformat the enshrined terminologies of psychological terms toward semantic frameworks of evidence-based mental categories, rather than perpetuating reified legacy terms from a previous historical era. Emerging LLM technologies can spark advances toward a biologically grounded redefinition of major brain disease nosology, cutting across diagnostic boundaries in a new era of evidence-based psychiatry, rather than relying on the judgment of selected experts alone. As Wittgenstein said, “The limits of my language mean the limits of my world.”⁶⁸

CONCLUSIONS

Biology has become “computable” over the last 5–10 years, such as in the form of massive genetic databases combined with targeted CRISPR gene editing and machine learning analytics, bringing us a step closer to an engineering discipline. Our demonstrated ability to generate biomolecular data troves eclipses our realized ambition to actually glean understanding from these systems—neuroscientists today are literally “drowning in information but starving for knowledge” (as written by John Naisbitt).⁶⁹

LLMs offer a new opportunity set in the game (but see Box 1). This model class shows that sheer statistical brute force can assist in demystifying the brain and disease by reading and generating biology, by crafting knowledge frameworks, and by unlocking never-before accessible modes of information integration and interrogation at scale. Foundation models will probably serve to extract, synergize, and synthesize knowledge from and across neuroscience domains, siloed “bubbles,” a task that may or may not exceed human comprehension. Neuroscientists will need to open up to and embrace the uncomfortable possibility that the human brain is a biological system that goes beyond what human intelligence alone can fully grasp without the assistance of AI tools applied to big data.

From a broader societal perspective, the industrial revolution touched mostly blue-collar jobs. In contrast, the current LLM revolution will perhaps mostly touch white-collar jobs, including

those of research workers in the neurosciences. Indeed, the unreasonable effectiveness of LLMs has been compared by venture capitalists and investors to the invention of fire as a tool, electricity, or the internet.

ACKNOWLEDGMENTS

D.B. was supported by the Brain Canada Foundation, through the Canada Brain Research Fund, with the financial support of Health Canada, National Institutes of Health (NIH R01 AG068563A, NIH R01 DA053301-01A1, NIH R01 MH129858-01A1), the Canadian Institute of Health Research (CIHR 438531, CIHR 470425), the Healthy Brains Healthy Lives initiative (Canada First Research Excellence fund), Google (Research Award, Teaching Award), and by the CIFAR Artificial Intelligence Chairs program (Canada Institute for Advanced Research).

DECLARATION OF INTERESTS

Four co-authors are employees at MindState Design Labs (A.T., O.L., P.W., and T.R.) and five are equity holders (D.B., A.T., O.L., P.W., and T.R.).

REFERENCES

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). Distributed representations of words and phrases and their compositional-ity. *Adv. Neural Inf. Process. Syst.* 26. https://papers.nips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.
- Le, Q., and Mikolov, T. (2014). Distributed representations of sentences and documents. *PMLR* 32, 1188–1196.
- Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1705.02364>.
- McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. *Adv. Neural Inf. Process. Syst.* <https://dl.acm.org/doi/10.5555/3295222.3295377>.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1301.3781>.
- Pennington, J., Socher, R., and Manning, C.D. (2014). Glove: Global vectors for word representation. <https://nlp.stanford.edu/pubs/glove.pdf>.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., and Lundberg, S. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2303.12712>.
- Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nas-tase, S.A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* 25, 369–380. <https://doi.org/10.1038/s41593-022-01026-4>.
- Caucheteux, C., Gramfort, A., and King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* 7, 430–441. <https://doi.org/10.1038/s41562-022-01516-2>.
- Schrimpf, M., Blank, I.A., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J.B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. USA* 118, e2105646118. <https://doi.org/10.1073/pnas.2105646118>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.
- Hassid, M., Peng, H., Rotem, D., Kasai, J., Montero, I., Smith, N.A., and Schwartz, R. (2022). How much does attention actually attend? Questioning the Importance of Attention in Pretrained Transformers. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2211.03495>.

13. Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. (2020). Long range arena: A benchmark for efficient transformers. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2011.04006>.
14. Bzdok, Danilo, and Yeo, B.T.T (2017). Inference in the age of big data: Future perspectives on neuroscience. *Neuroimage* 155, 549–564.
15. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., and Metzler, D. (2022). Emergent abilities of large language models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2206.07682>.
16. OpenAI. (2023). GPT-4 Technical Report. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2303.08774>.
17. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2001.08361>.
18. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., and Azhar, F. (2023). Llama: Open and efficient foundation language models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2302.13971>.
19. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., and Clark, A. (2022). Training compute-optimal large language models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2203.15556>.
20. Schaeffer, R., Miranda, B., and Koyejo, S. (2023). Are emergent abilities of Large Language Models a mirage?. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2304.15004>.
21. Caballero, E., Gupta, K., Rish, I., and Krueger, D. (2022). Broken neural scaling laws. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2210.14891>.
22. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. *PMLR* 97, 2790–2799. <https://proceedings.mlr.press/v97/houlsby19a/houlsby19a.pdf>.
23. Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., and Gurevych, I. (2020). Adapterhub: A framework for adapting transformers. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2007.07779>.
24. Bapna, A., Arivazhagan, N., and Firat, O. (2019). Simple, scalable adaptation for neural machine translation. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1909.08478>.
25. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog* 1, 9.
26. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901.
27. Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., and Hu, Z. (2023). Language Models Meet World Models: Embodied Experiences Enhance Language Models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2305.10626>.
28. Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T., and Evans, O. (2023). The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". Preprint at arXiv. <https://doi.org/10.48550/arXiv.2309.12288>.
29. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J., and Ntranos, V. (2023). Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* 55, 1512–1522. <https://doi.org/10.1038/s41588-023-01465-0>.
30. Cui, H., Wang, C., Maan, H., and Wang, B. (2023). scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI. Preprint at bioRxiv. <https://doi.org/10.1101/2023.04.30.538439>.
31. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
32. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA* 118, e2016239118. <https://doi.org/10.1073/pnas.2016239118>.
33. Yang, E., Milisav, F., Kopal, J., Holmes, A.J., Mitsis, G.D., Misic, B., Finn, E.S., and Bzdok, D. (2023). The default network dominates neural responses to evolving movie stories. *Nat. Commun.* 14, 4197. <https://doi.org/10.1038/s41467-023-39862-y>.
34. Ye, Z., Liu, Y., and Li, Q. (2021). Recent Progress in Smart Electronic Nose Technologies Enabled with Machine Learning Methods. *Sensors* 21, 7620. <https://doi.org/10.3390/s21227620>.
35. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., and Reynolds, M. (2022). Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* 35, 23716–23736.
36. Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. <https://aclanthology.org/P18-1238/>.
37. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 64–73.
38. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al. (2023). A foundation model for generalizable disease detection from retinal images. *Nature* 622, 156–163.
39. Wagner, S.K., Hughes, F., Cortina-Borja, M., Pontikos, N., Struyven, R., Liu, X., Montgomery, H., Alexander, D.C., Topol, E., Petersen, S.E., et al. (2022). AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353 157 patients in London, UK. *BMJ open* 12, e058552.
40. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36.
41. Bzdok, D., and Ioannidis, J.P. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends in neurosciences* 42, 251–262.
42. Bzdok, D., Engemann, D., and Thirion, B. (2020). Inference and prediction diverge in biomedicine. *Patterns* 1, 100119.
43. Shanahan, M., McDonnell, K., and Reynolds, L. (2023). Role play with large language models. *Nature* 623, 493–498. <https://doi.org/10.1038/s41586-023-06647-8>.
44. Sharma, A., Kumar, R., Ranjta, S., and Varadwaj, P.K. (2021). SMILES to smell: decoding the structure–odor relationship of chemical compounds using the deep neural network approach. *J. Chem. Inf. Model.* 61, 676–688.
45. Ballentine, G., Friedman, S.F., and Bzdok, D. (2022). Trips and neurotransmitters: Discovering principled patterns across 6850 hallucinogenic experiences. *Sci. Adv.* 8, eabl6989.
46. Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. (2023). Pmc-llama: Further finetuning llama on medical papers. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2304.14454>.
47. Rodziewicz, T.L., Houseman, B., and Hipskind, J.E. (2023). Medical Error Reduction and Prevention. In *StatPearls* (StatPearls Publishing LLC).
48. Hipp, R., Abel, E., and Weber, R.J. (2016). A Primer on Clinical Pathways. *Hosp. Pharm.* 51, 416–421. <https://doi.org/10.1310/hpj5105-416>.
49. Acosta, J.N., Falcone, G.J., Rajpurkar, P., and Topol, E.J. (2022). Multimodal biomedical AI. *Nat. Med.* 28, 1773–1784. <https://doi.org/10.1038/s41591-022-01981-2>.

50. Harter, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBio-Medicine* 90, 104512. <https://doi.org/10.1016/j.ebiom.2023.104512>.
51. Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., and Gallant, J.L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron* 63, 902–915.
52. Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* 8, 665–670. <https://doi.org/10.1038/nmeth.1635>.
53. Laird, A.R., Lancaster, J.L., and Fox, P.T. (2005). BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics* 3, 65–78. <https://doi.org/10.1007/s10238-005-0011-1> [pii].
54. Fox, P.T., and Lancaster, J.L. (2002). Opinion: Mapping context and content: the BrainMap model. *Nat. Rev. Neurosci.* 3, 319–321. [pii]. <https://doi.org/10.1038/nrn789>.
55. Krakauer, J.W., Ghazanfar, A.A., Gomez-Marín, A., MacIver, M.A., and Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron* 93, 480–490. <https://doi.org/10.1016/j.neuron.2016.12.041>.
56. Pessoa, L. (2008). On the relationship between emotion and cognition. *Nat. Rev. Neurosci.* 9, 148–158. <https://doi.org/10.1038/nrn2317>.
57. Van Overwalle, F. (2011). A dissociation between social mentalizing and general reasoning. *Neuroimage* 54, 1589–1599. <https://doi.org/10.1016/j.neuroimage.2010.09.043> [pii].
58. Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A.R., Langner, R., and Eickhoff, S.B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Struct. Funct.* 217, 783–796. <https://doi.org/10.1007/s00429-012-0380-y>.
59. Dohmatob, E., Dumas, G., and Bzdok, D. (2020). Dark control: The default mode network as a reinforcement learning agent. *Hum. Brain Mapp.* 41, 3318–3341. <https://doi.org/10.1002/hbm.25019>.
60. Spreng, R.N., Mar, R.A., and Kim, A.S.N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *J. Cogn. Neurosci.* 21, 489–510. <https://doi.org/10.1162/jocn.2008.21029>.
61. György Buzsáki, M. (2019). *The brain from inside out* (Oxford University Press).
62. Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* 10, 59–63. <https://doi.org/10.1016/j.tics.2005.12.004> [pii].
63. Laird, A.R., Fox, P.M., Eickhoff, S.B., Turner, J.A., Ray, K.L., McKay, D.R., Glahn, D.C., Beckmann, C.F., Smith, S.M., and Fox, P.T. (2011). Behavioral interpretations of intrinsic connectivity networks. *J. Cogn. Neurosci.* 23, 4022–4037. https://doi.org/10.1162/jocn_a_00077.
64. Mesulam, M.M. (1998). From sensation to cognition. *Brain* 121 (Pt 6), 1013–1052.
65. Voytek, B. (2022). The data science future of neuroscience theory. *Nat. Methods* 19, 1349–1350. <https://doi.org/10.1038/s41592-022-01630-z>.
66. Brainstorm Consortium, Anttila, V., Bulik-Sullivan, B., Finucane, H.K., Walters, R.K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G.J., Gormley, P., et al. (2018). Analysis of shared heritability in common disorders of the brain. *Science* 360, eaap8757. <https://doi.org/10.1126/science.aap8757>.
67. Beam, E., Potts, C., Poldrack, R.A., and Etkin, A. (2021). A data-driven framework for mapping domains of human neurobiology. *Nat. Neurosci.* 24, 1733–1744. <https://doi.org/10.1038/s41593-021-00948-9>.
68. Wittgenstein, L. (1958). *Philosophical Investigations* (Basil Blackwell).
69. Naisbitt, J. (1988). *Megatrends: ten new directions transforming our lives* (Warner Books).
70. Dziri, N., Milton, S., Yu, M., Zaiane, O., and Reddy, S. (2022). On the origin of hallucinations in conversational models: Is it the datasets or the models?. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2204.07931>.
71. Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1906.02243>.
72. Nadeem, M., Bethke, A., and Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2004.09456>.
73. Liu, F., Bugliarello, E., Ponti, E.M., Reddy, S., Collier, N., and Elliott, D. (2021). Visually grounded reasoning across languages and cultures. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2109.13238>.