Welcome!

https://forms.gle/eUiwf2ENxxk6d9ig8

Before we begin, please complete this short survey

Natural language processing with LLMs in psychology research

PBHM Workshop, 4 February 2026

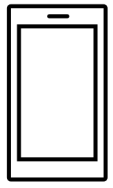
Judith Mildner, PhD

Gillan Lab

Why this workshop?

Language offers unique access to people's mental lives

I use language to study real-world cognition and mental health



Language and depressive symptoms in mental health app

Contents and dynamics of the stream of thought

Why this workshop?

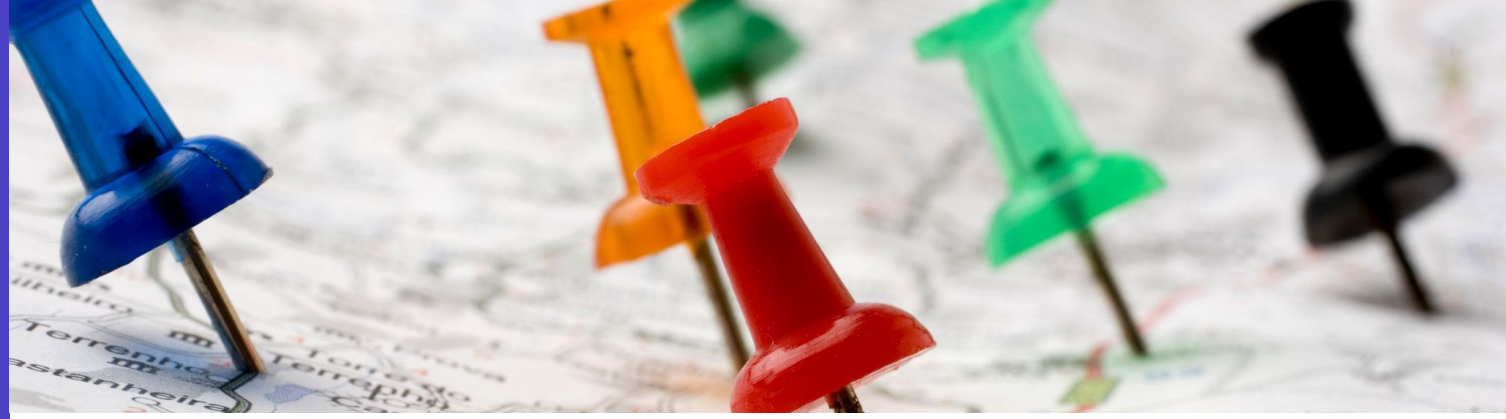
Natural language processing is booming

Analysis techniques are becoming more powerful and more accessible

Changes are happening quickly, lots of new AI-related terminology and buzzwords

If you want to do NLP, where do you even start?

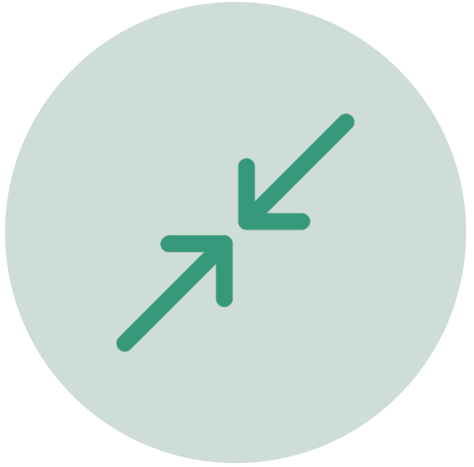
Agenda



- Introduction to natural language processing (NLP)
- Transformer models and key technical concepts
- Brief NLP demo
- You try!

Why language?

RICH BEHAVIORAL DATA,
POSSIBLE TO COLLECT IN NATURAL CONTEXTS



SCALABLE
LOW COST AND ACCESSIBLE

NLP in psychology

- Language as data in psychology is not new
- Several influential methodological developments in recent history
  - Bag of words
  - Word embeddings
  - Transformer models

Bag of words

Analyzing all words, regardless of their order

Most popular method: Word counting programs (Pennebaker & King, 1999)

1. Develop dictionaries (and validate them)
2. Count words in text matching each dictionary
3. Profit!

Linguistic Inquiry and Word Count (LIWC) software (Pennebaker, 2003) has dominated natural language processing in psychology in the last few decades

Weakness: language is ambiguous, impossible to categorize every word

Word embeddings

Lists of numbers that represent meaning of words

Capturing semantic and syntactic relationships between words, e.g. Word2vec (Mikolov et al, 2013) and GloVe (Pennington et al, 2014)

This technique unlocked measures such as semantic similarity

Weakness: in original models, each word only has 1 embedding. Not context-dependent

Words in context

"I crossed the river to get to the other **bank**"

"I crossed the road to get to the other **bank**"

Humans:

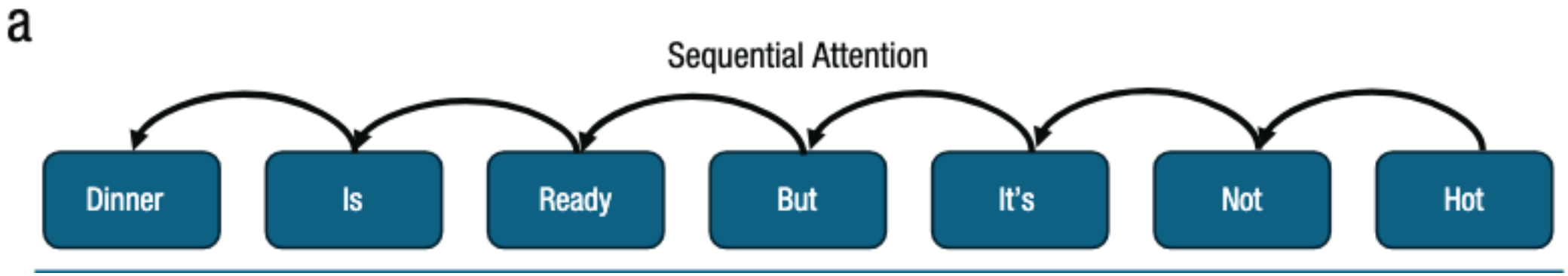
Bag of words:

Word embedding:

Attention

Using deep learning (e.g. recurrent neural nets, long short-term memory networks) to update word embeddings based on surrounding words

Dynamic algorithm instead of set rules for each word



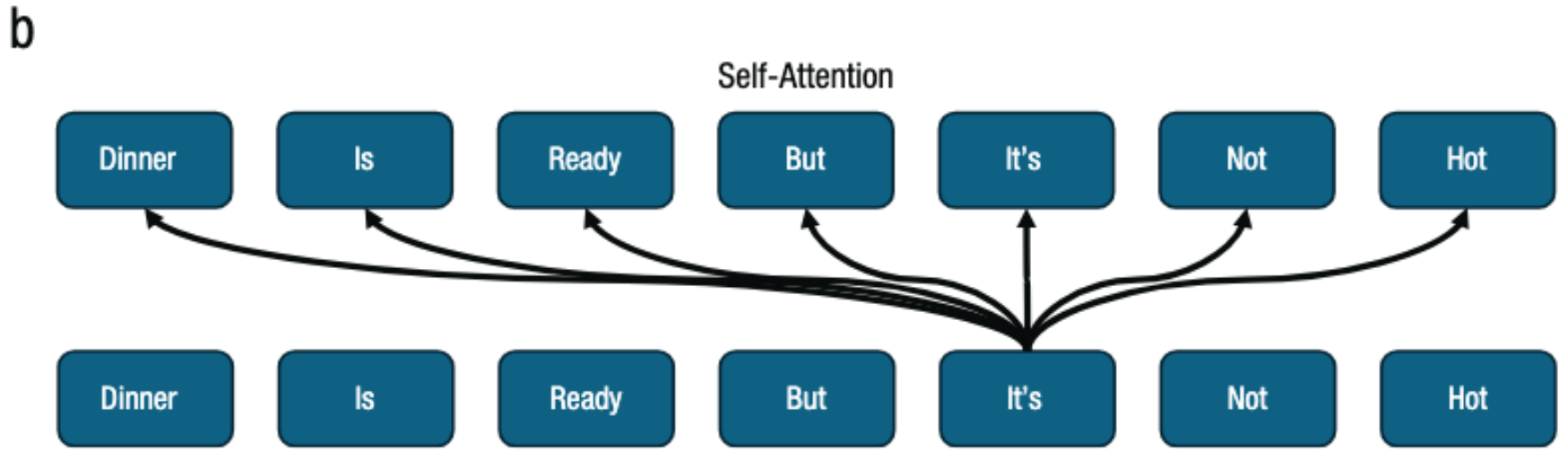
Brickman et al., 2025

Attention is all you need

Vaswani et al., 2017

Introduced transformer models

Foundation for modern language models, including LLMs



Brickman et al., 2025

Transformer model

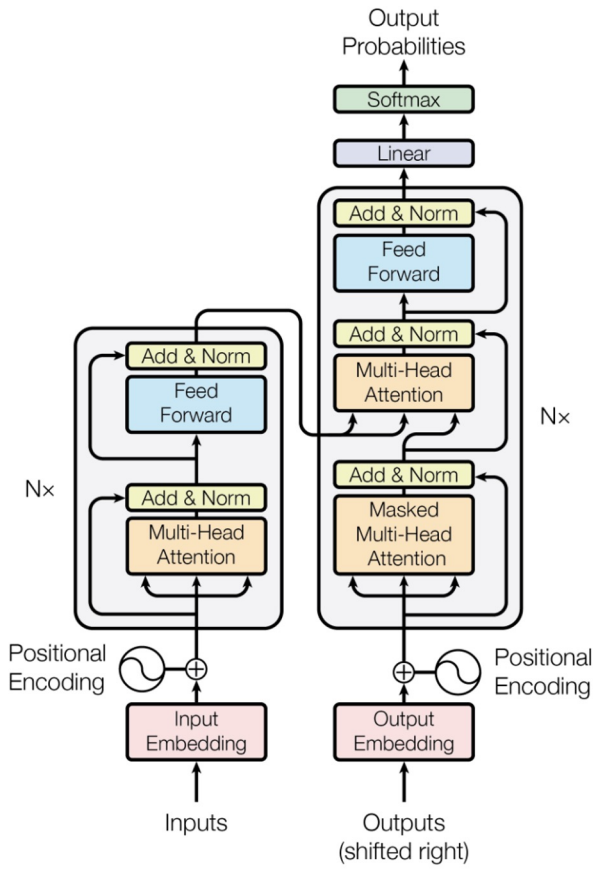


Figure 1: The Transformer - model architecture.

Vaswani et al., 2017

Transformers consist of encoders and decoders with self-attention layers

Encoder:
Learns relationships among input tokens

Decoder:
Can process output from encoder and generate text

Transformer architecture can handle large amounts of data efficiently and is sensitive to context

Model size (# of model parameters) grew from millions to billions → large language models (LLMs)

Transformer types

Encoder only

Good at understanding input texts

Example tasks: classification, sentiment analysis, NER

Famous models: BERT, RoBERTa, DeBERTa, etc

Decoder only

Good at generative tasks, responses predicted one token at a time

Example tasks: chatbots, text generation

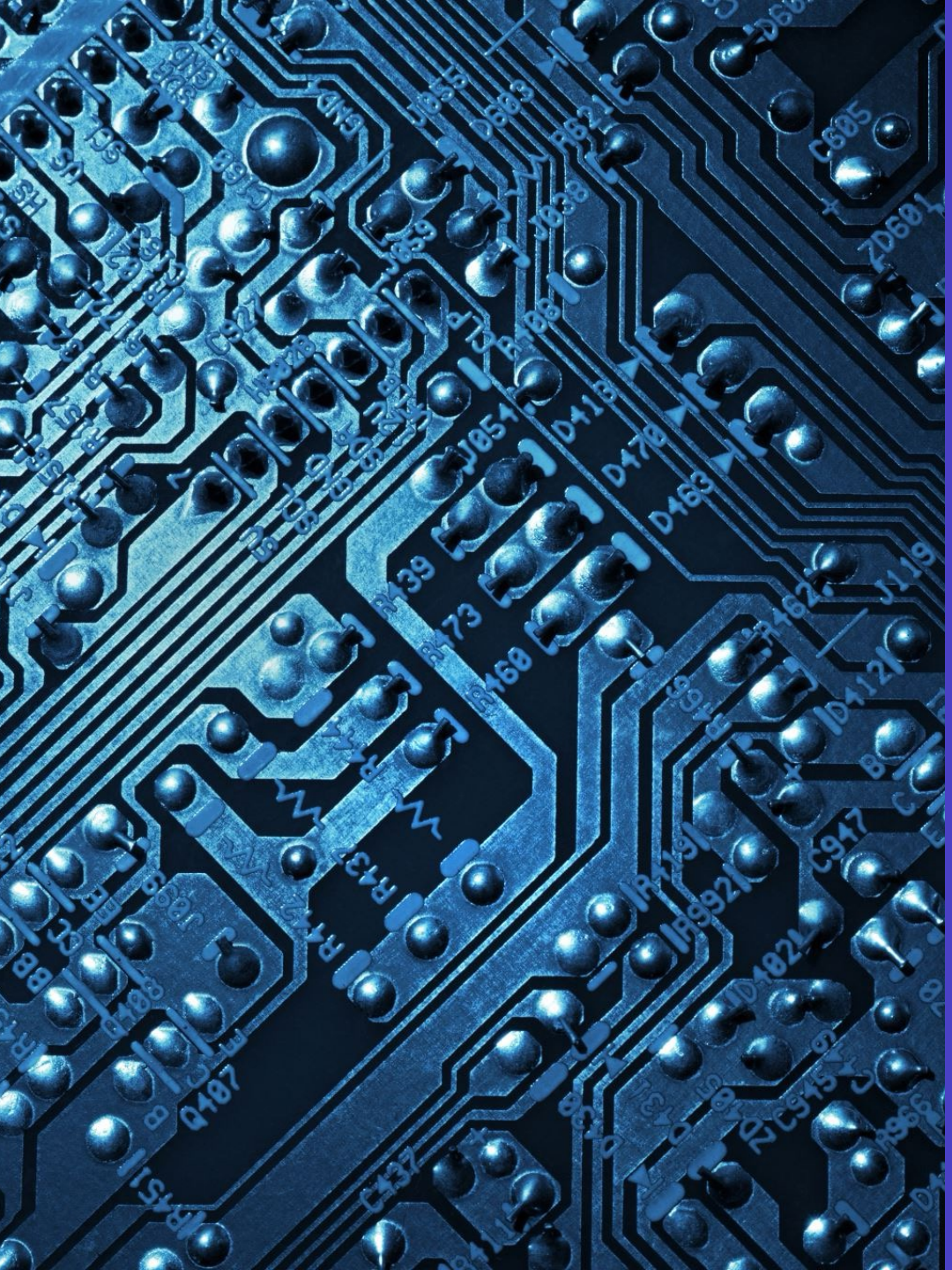
Famous models: GPT, Llama, etc

Encoder-Decoder

Good when need to produce output that is dependent on understanding of input

Example tasks: translation, speech to text, summarization

Famous models: T5, BART, WhisperAI



Using transformer models in psychology research

Key technical concepts

Workflow - preparation

1: Data Collection	2: Language Conversion	3: Text Preprocessing
Choose source:	Change data format to text	Text Isolation
Interview Language	Audio Processing	De-identification
Social Media	Other Conversion	Tokenization
Spoken or written prompts		
Electronic Health Records		

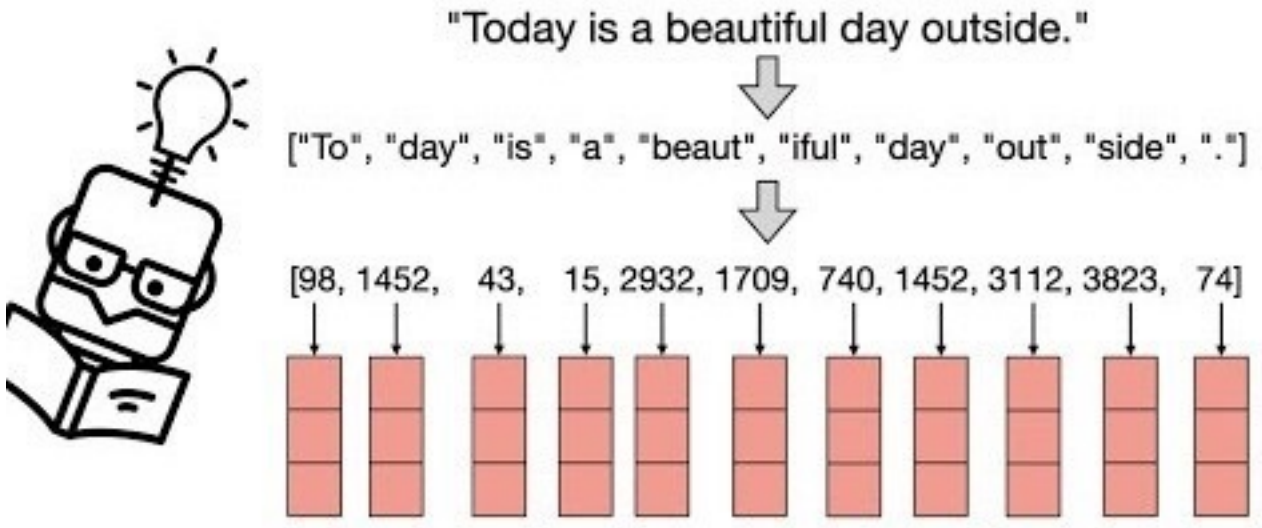
Adapted from Brickman J, Gupta M, Oltmanns JR. Large Language Models for Psychological Assessment: A Comprehensive Overview. Advances in Methods and Practices in Psychological Science. 2025;8(3). doi:10.1177/25152459251343582

Tokenization

Process of converting words to numeric input for model

Most LLMs have their own tokenizer

Typically 1 token per 0.75 words as some words get broken down into subunits, and there are special tokens such as sentence separation.



https://www.linkedin.com/pulse/tokenization-how-llms-process-text-tokens-nikitha-r-gnbkf

Workflow – LLM stage

TODAY

4: LLM Technique	5: LLM Selection	6: Model Evaluation	7: Model Training Considerations	8: Model Visualization
Feature Extraction Fine Tuning Prompt Engineering Ex: Zero-shot prompting	Selection based on: Data characteristics Sample size Computing power	Prediction Ex: regression vs. classification Performance methods	Cross validation Hyperparameter tuning	Topic Modeling Heatmaps Embedding visualization

Adapted from Brickman J, Gupta M, Oltmanns JR. Large Language Models for Psychological Assessment: A Comprehensive Overview. Advances in Methods and Practices in Psychological Science. 2025;8(3). doi:10.1177/25152459251343582

How to implement LLMs

API Calls – models in the cloud

Model providers allow users to send requests to their models

You send data and payment, they run data through their model and return the result

Pros:
- Easy, quick, usually cheap
- No special hardware required
- State of the art models

Drawbacks:
- Lack of privacy/transparency
- Hard to reproduce (model versions change without notice)

Local usage

Download a model yourself and run it on your data locally

Pros:
- Private, suitable for sensitive data
- Replicable, as you can keep the exact same model

Drawbacks:
- Requires more technical expertise
- Requires access to GPUs
- Model size and versions limited by hardware and open-source availability

Questions?

Next up: code demo

Software

Many options available for local LLMs

Llama.cpp

Ollama

HuggingFace

vLLM

Hugging Face

Hub for open-source NLP tech

Known for 3 things:

- Code libraries, most importantly the Transformers library
- Model hub: massive repository of pretrained models
- Open datasets: lots of machine learning datasets in different modalities, domains, and languages

Demo