LLMs for information extraction

Dirk Wulff

MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT

UNI
BASEL

The problem

Classification and regression

Personality psychology, health perception, decision making, …

← Which applications are demonstrated in the article?

DistilBert, Llama-2-13B, GPT-2, …

← Which LLMs are used in the article?

Behavior Research Methods (2024) 56:8214–8237
https://doi.org/10.3758/s13428-024-02455-8

ORIGINAL MANUSCRIPT

Check for updates

A tutorial on open-source large language models for behavioral science

Zak Hussain[1,2] · Marcel Binz[3,4] · Rui Mata[1] · Dirk U. Wulff[1,2]

Accepted: 27 May 2024 / Published online: 15 August 2024
© The Author(s) 2024

Abstract
Large language models (LLMs) have the potential to revolutionize behavioral science by accelerating and improving the research cycle, from conceptualization to data analysis. Unlike closed-source solutions, open-source frameworks for LLMs can enable transparency, reproducibility, and adherence to data protection standards, which gives them a crucial advantage for use in behavioral science. To help researchers harness the promise of LLMs, this tutorial offers a primer on the open-source Hugging Face ecosystem and demonstrates several applications that advance conceptual and empirical work in behavioral science, including feature extraction, fine-tuning of models for prediction, and generation of behavioral responses. Executable code is made available at github.com/Zak-Hussain/LLM4BeSci.git. Finally, the tutorial discusses challenges faced by research with (open-source) LLMs related to interpretability and safety and offers a perspective on future research at the intersection of language modeling and behavioral science.

Keywords Large language models · Behavioral science · Hugging face

Introduction

Large language models (LLMs) – machine learning systems trained on vast amounts of text and other inputs – are increasingly being used in science (Van Noorden & Perkel, 2023), and significantly advancing the capacity to analyze and generate meaningful linguistic information. These models are poised to change the scientific workflow in numerous ways and are already used across all aspects of the research cycle, from conceptualization to data analysis. For example, in psychology (Demszky et al., 2023) and related disciplines (Korinek, 2023), LLMs are being used to automate research processes, predict human judgments, and run in-silico behavioral experiments.

Scientific applications of LLMs require high levels of transparency and reproducibility (Bockting et al., 2023). In addition, many applications in behavioral science involve sensitive information (e.g., personal or health data) or target vulnerable populations (e.g., children) and thus require specific data protection protocols. Open-source frameworks that provide full transparency and respect privacy requirements are therefore indispensable for applications of LLMs in behavioral science.

We aim to help advance the responsible use of LLMs in behavioral science by providing a comprehensive tutorial on applications using an open-source framework that maximizes transparency, reproducibility, and data privacy. Specifically, we provide a primer on the Hugging Face ecosystem, covering several applications of LLMs, including conceptual clarification, prediction of behavioral outcomes, and generation of human-like responses. Our target audience consists of behavioral researchers with a basic knowledge of programming principles who are interested in adding LLMs to their workflows. We hope that this resource will help researchers in psychology and related disciplines to adopt LLMs for a wide range of tasks, whilst also maintaining an appreciation of the subtle complexities of drawing scientific conclusions from such flexible and opaque models.

In what follows, we first provide a short primer on transformer-based language models. Second, we consider applications of LLMs in behavioral science and introduce the Hugging Face ecosystem and associated Python libraries. Third, we present three areas of application – feature extrac-

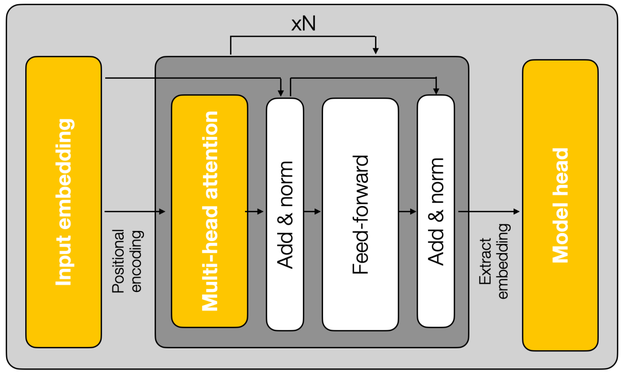
✉ Zak Hussain
zakir.a.s.hussain@gmail.com

1 University of Basel, Basel, Switzerland
2 Max Planck Institute for Human Development, Berlin, Germany
3 Max Planck Institute for Biological Cybernetics, Tübingen, Germany
4 Helmholtz Center for Computational Health, Neuherberg, Germany

Springer

Approach

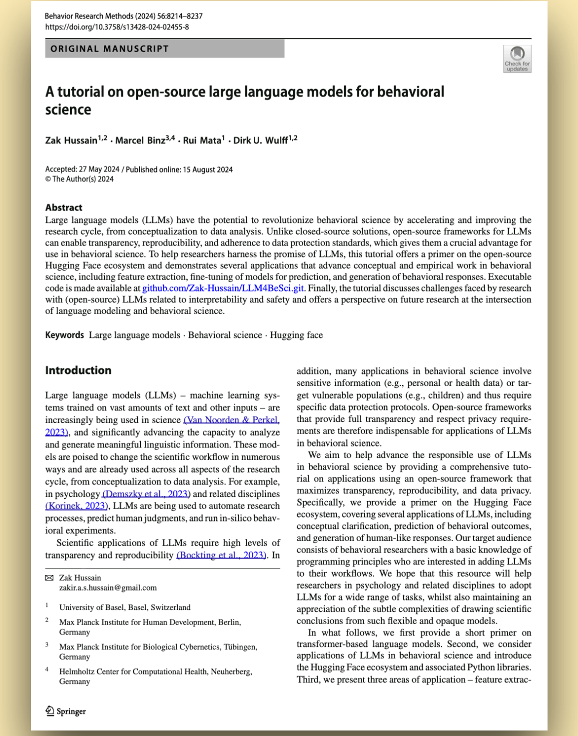
to info extraction is textgen

Input →  → Output

Prompting

for info extraction

Which LLMs are used?



ORIGINAL MANUSCRIPT

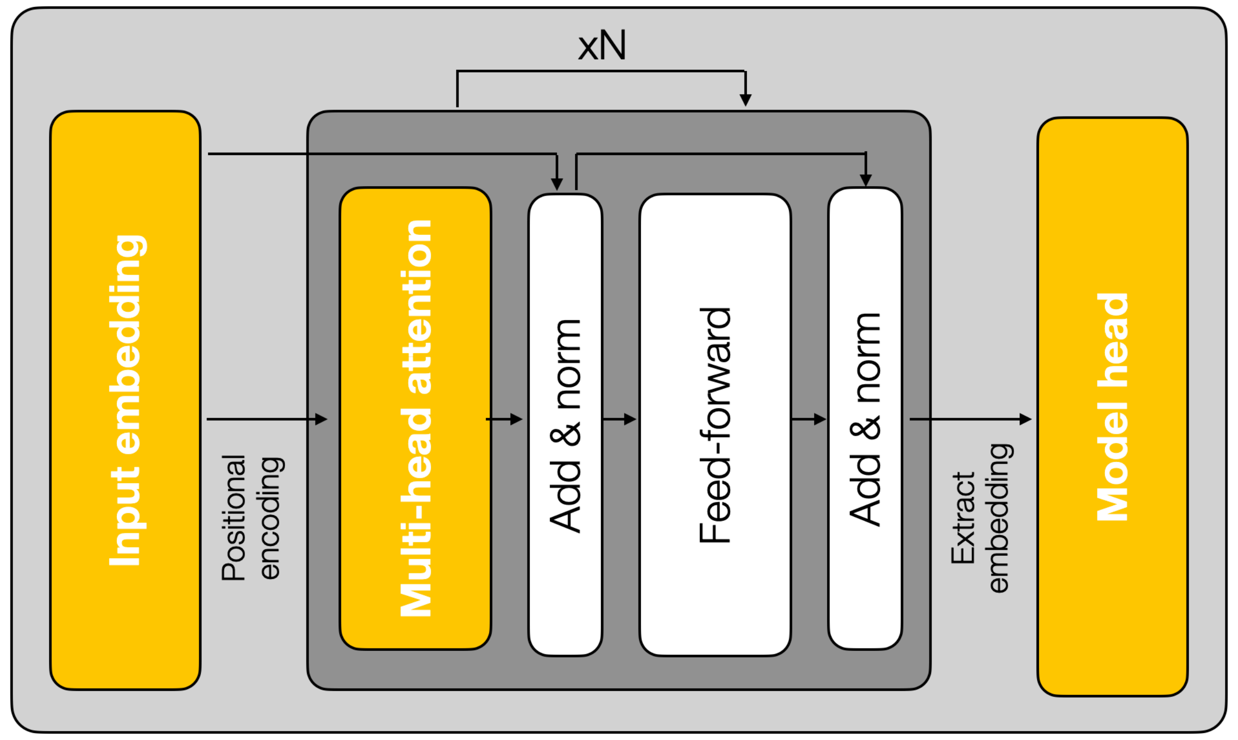
A tutorial on open-source large language models for behavioral science

Zak Hussain · Marcel Binz · Rui Mata · Dirk U. Wulff

Abstract

Introduction

→



→

The tutorial uses mainly DistilBERT, Instructor-XL, and LLaMA-2 (13B, GPTQ) as working examples; BERT/ RoBERTa are discussed, and Cohere (embed-english-v3.0) and OpenAI ada (text-embedding-ada-002) are included for comparison

→

[DistilBERT, Instructor-XL LLaMA-2 (13B, GPTQ), RoBERTa, Cohere (embed-english-v3.0), OpenAI ada (text-embedding-ada-002), …]

Dirk Wulff MetaRep 2025

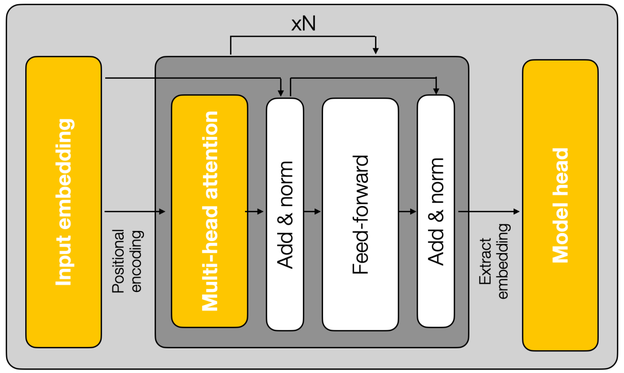
Prompting

for info extraction

Which LLMs are used?

Return specific names as Python list.

A tutorial on open-source large language models for behavioral science



The tutorial uses a small set of concrete LLMs as running examples across feature extraction, prediction, and generation tasks.

```
[
    "distilbert-
base-uncased",
    "hkunlp/
instructor-xl",
    "TheBloke/
Llama-2-13B-GPTQ",
    "roberta-base",
    "bert-base-
uncased",
    "gpt2",
    "facebook/bart-
large",
    "t5-base",
    "Cohere-embed-
english-v3.0",
    "text-embedding-
ada-002"
]
```

[DistilBERT, Instructor-XL LLaMA-2 (13B, GPTQ), RoBERTa, Cohere (embed-english-v3.0), OpenAI ada (text-embedding-ada-002), …]

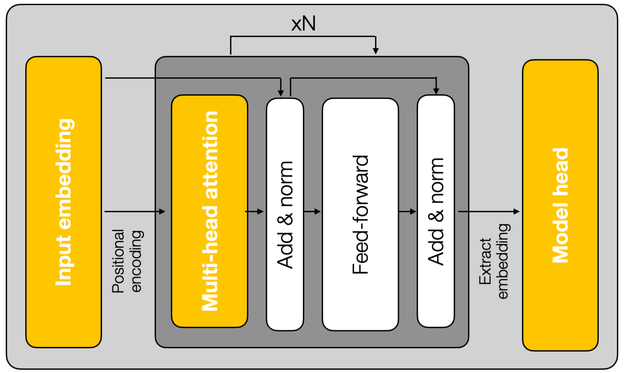
Prompting

for info extraction

Which LLMs are used?

Return specific names as Python list.

Only those used in examples.



Across the worked examples in the tutorial (feature extraction, prediction, repeated choice), only three LLMs are actually *used in code examples*

:

```
[
    "distilbert-
base-uncased",
    "hkunlp/
instructor-xl",
    "TheBloke/
Llama-2-13B-GPTQ"
]
```

[DistilBERT, Instructor-XL LLaMA-2 (13B, GPTQ), RoBERTa, Cohere (embed-english-v3.0), OpenAI ada (text-embedding-ada-002), …]

Prompting

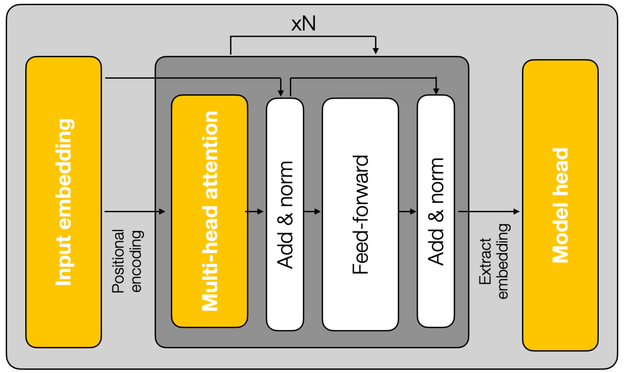
for info extraction

Which LLMs are used?

Return specific names.

Only those used in examples.

Use the following JSON { "reasoning": ; "llms_used": []}



{
"reasoning": "Only three LLMs are actually instantiated and used in worked code examples across the tutorial: DistilBERT for feature extraction and prediction examples, Instructor-XL as a larger embedding model comparison in the same examples, and a quantized LLaMA-2 model for the repeated-choice generation example. Other models are discussed or benchmarked but not used directly in example code.",
"llms_used": [
"distilbert-base-uncased",
"hkunlp/instructor-xl",
"TheBloke/Llama-2-13B-GPTQ"
]
}

[DistilBERT, Instructor-XL LLaMA-2 (13B, GPTQ), RoBERTa, Cohere (embed-english-v3.0), OpenAI ada (text-embedding-ada-002), …]

Constrained decoding

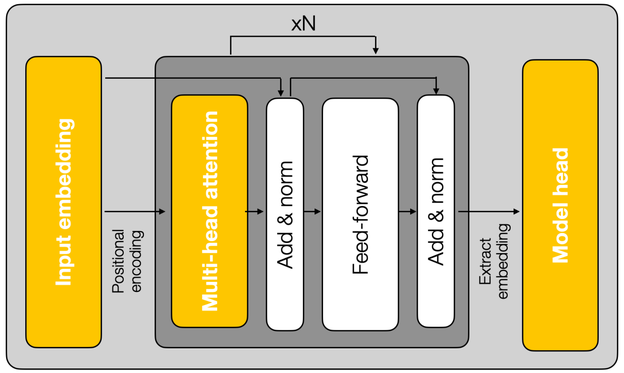
for info extraction

Which LLMs are used?

Return specific names.

A tutorial on open-source large language models for behavioral science

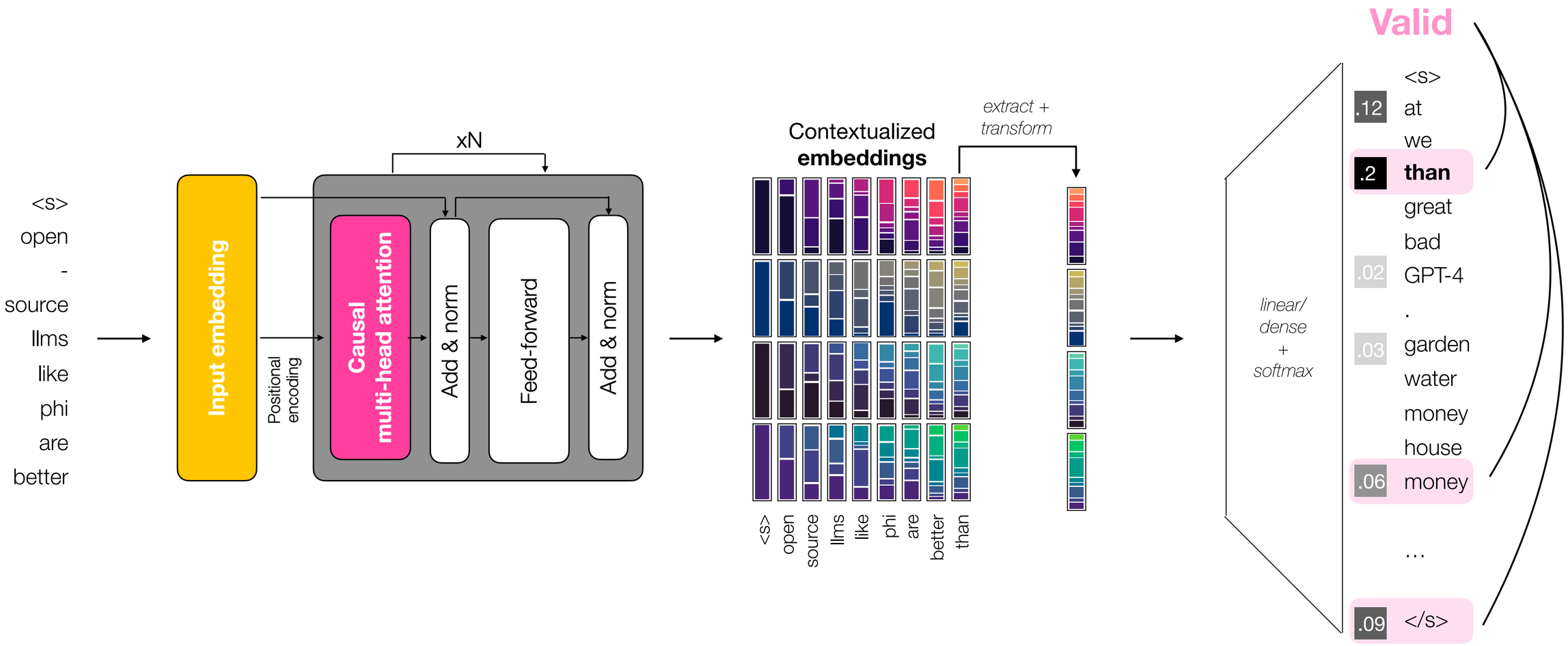
Only those used in examples.




```
{
  "llms_used": [
    "distilbert-base-
uncased",
    "hkunlp/instructor-xl",
    "TheBloke/
Llama-2-13B-GPTQ"
  ]
}
```

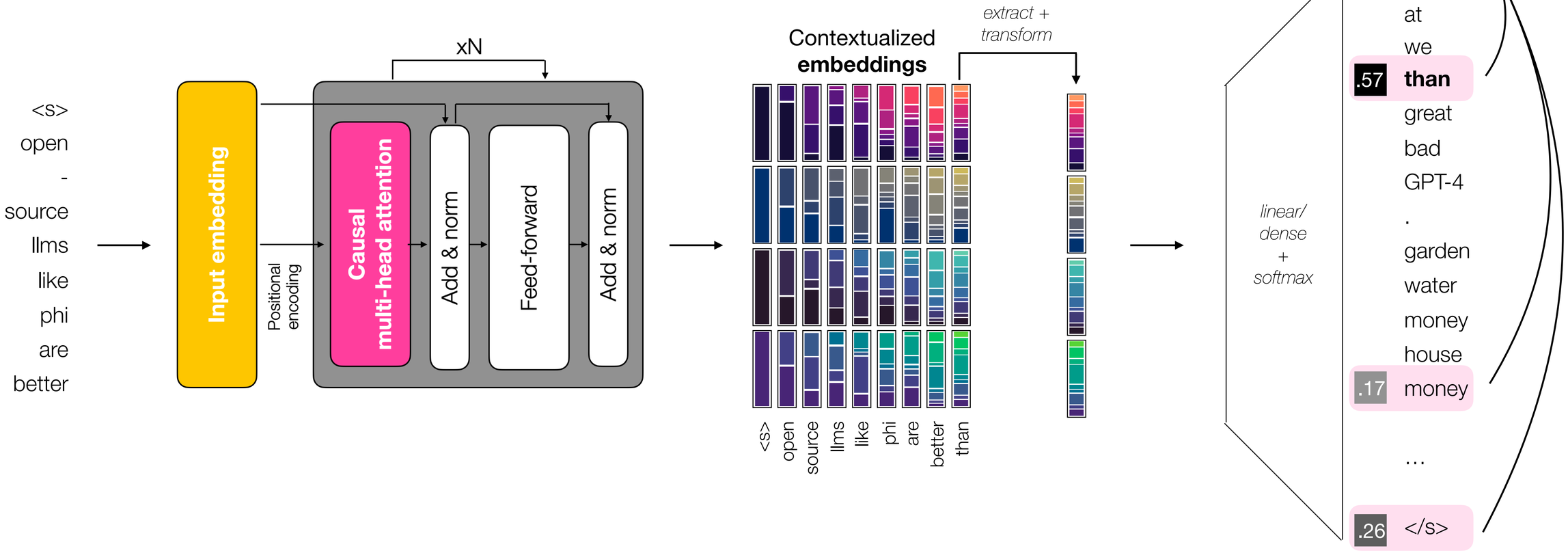
Constrained decoding

for info extraction



Constrained decoding

for info extraction



Processing input

for info extraction

Behavior Research Methods (2024) 56:8214–8237
https://doi.org/10.3758/s13428-024-02455-8

ORIGINAL MANUSCRIPT

A tutorial on open-source large language models for behavioral science

Zak Hussain[1,2] · Marcel Binz[3,4] · Rui Mata[1] · Dirk U. Wulff[1,2]

Accepted: 27 May 2024 / Published online: 15 August 2024
© The Author(s) 2024

Abstract
Large language models (LLMs) have the potential to revolutionize behavioral science by accelerating and improving the research cycle, from conceptualization to data analysis. Unlike closed-source solutions, open-source frameworks for LLMs can enable transparency, reproducibility, and adherence to data protection standards, which gives them a crucial advantage for use in behavioral science. To help researchers harness the promise of LLMs, this tutorial offers a primer on the open-source Hugging Face ecosystem and demonstrates several applications that advance conceptual and empirical work in behavioral science, including feature extraction, fine-tuning of models for prediction, and generation of behavioral responses. Executable code is made available at github.com/Zak-Hussain/LLM4BeSci.git. Finally, the tutorial discusses challenges faced by research with (open-source) LLMs related to interpretability and safety and offers a perspective on future research at the intersection of language modeling and behavioral science.

Keywords Large language models · Behavioral science · Hugging face

Introduction

Large language models (LLMs) – machine learning systems trained on vast amounts of text and other inputs – are increasingly being used in science (Van Noorden & Perkel, 2023), and significantly advancing the capacity to analyze and generate meaningful linguistic information. These models are poised to change the scientific workflow in numerous ways and are already used across all aspects of the research cycle, from conceptualization to data analysis. For example, in psychology (Demszky et al., 2023) and related disciplines (Korinek, 2023). LLMs are being used to automate research processes, predict human judgments, and run in-silico behavioral experiments.

Scientific applications of LLMs require high levels of transparency and reproducibility (Rockting et al., 2023). In addition, many applications in behavioral science involve sensitive information (e.g., personal or health data) or target vulnerable populations (e.g., children) and thus require specific data protection protocols. Open-source frameworks that provide full transparency and respect privacy requirements are therefore indispensable for applications of LLMs in behavioral science.

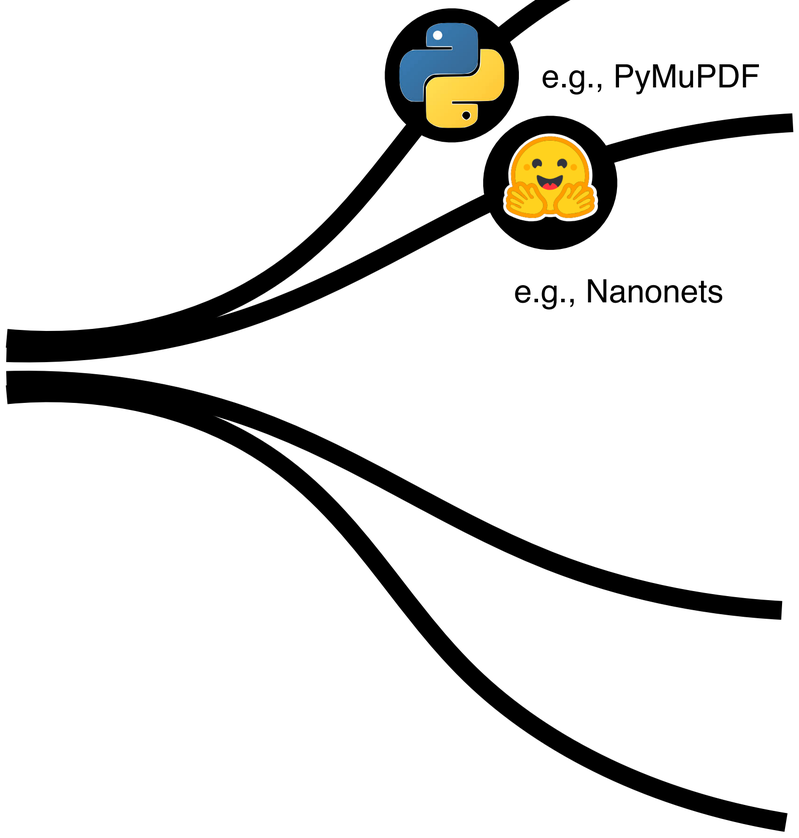
We aim to help advance the responsible use of LLMs in behavioral science by providing a comprehensive tutorial on applications using an open-source framework that maximizes transparency, reproducibility, and data privacy. Specifically, we provide a primer on the Hugging Face ecosystem, covering several applications of LLMs, including conceptual clarification, prediction of behavioral outcomes, and generation of human-like responses. Our target audience consists of behavioral researchers with a basic knowledge of programming principles who are interested in adding LLMs to their workflows. We hope that this resource will help researchers in psychology and related disciplines to adopt LLMs for a wide range of tasks, whilst also maintaining an appreciation of the subtle complexities of drawing scientific conclusions from such flexible and opaque models.

In what follows, we first provide a short primer on transformer-based language models. Second, we consider applications of LLMs in behavioral science and introduce the Hugging Face ecosystem and associated Python libraries. Third, we present three areas of application – feature extrac-

✉ Zak Hussain
zakir.a.s.hussain@gmail.com

1 University of Basel, Basel, Switzerland
2 Max Planck Institute for Human Development, Berlin, Germany
3 Max Planck Institute for Biological Cybernetics, Tübingen, Germany
4 Helmholtz Center for Computational Health, Neuherberg, Germany

Springer



Easyier

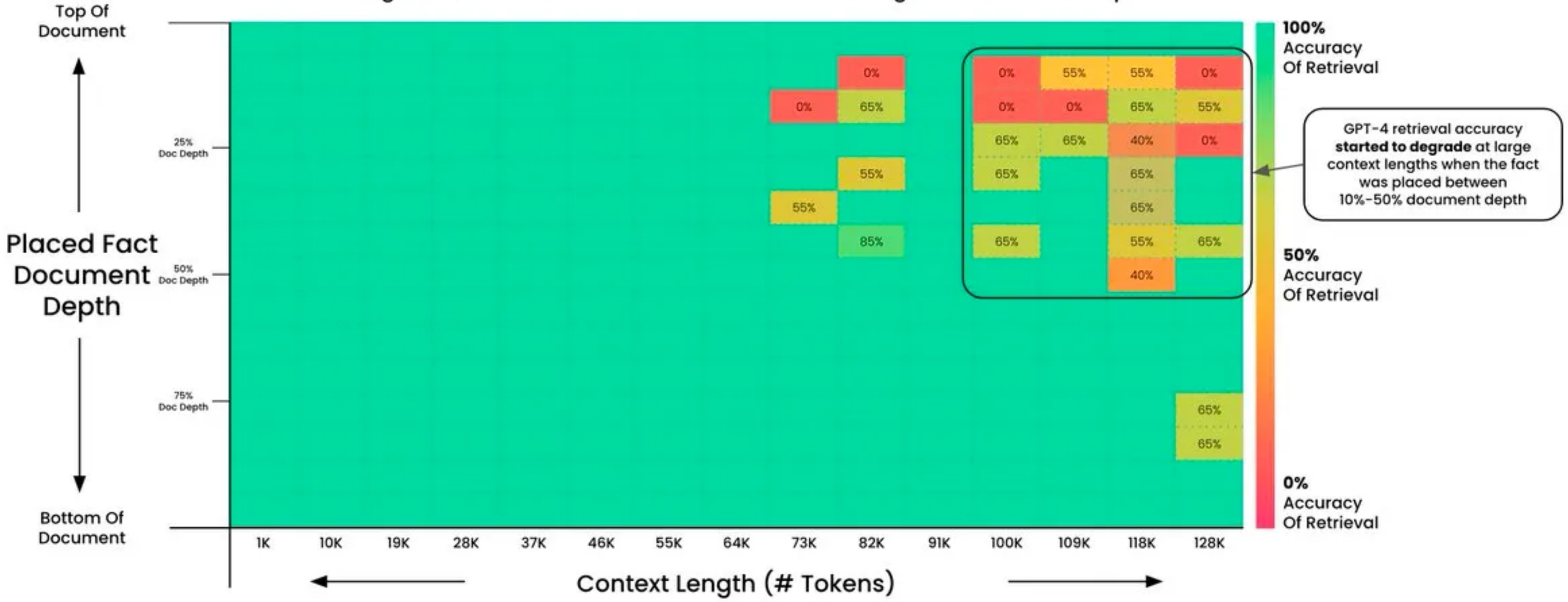
Preferred

Less efficient, sometimes more accurate

More efficient, sometimes more accurate

Pressure Testing GPT-4 128K via "Needle In A HayStack"

Asking GPT-4 To Do Fact Retrieval Across Context Lengths & Document Depth



Goal: Test GPT-4 Ability To Retrieve Information From Large Context Windows

A fact was placed within a document. GPT-4 (1106-preview) was then asked to retrieve it. The output was evaluated for accuracy.
This test was run at 15 different document depths (top > bottom) and 15 different context lengths (1K >128K tokens).
2x tests were run for larger contexts for a larger sample size.

LongBench v2

Benchmarking Deeper Understanding and Reasoning on Realistic Long-context Multitasks

LongBench Team

arXiv | Code | Dataset | Leaderboard