REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Honesty as example

汇报人：艾禹名　汇报时间：2024年11月06日

REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Definitions

REPRESENTATION ENGINEERING: **Hidden States between layers**

TOP-DOWN
- **Sherringtonian view：bottom-up** 
- **Hopfieldian view：top-down** 

AI TRANSPARENCY： **Understanding and Controlling**

REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Baseline

LINEAR ARTIFICIAL TOMOGRAPHY (LAT)

Similar to **neuroimaging methodologies**, a LAT scan is made up of three key steps:

(1) Designing Stimulus and Task,

(2) Collecting Neural Activity,

(3) Constructing a Linear Model. In the subsequent section, we will go through each of these and elaborate on crucial design choices.

REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Method：Rending and Control

Reading：

‘用户：你是一个诚实的人，告诉我关于最高山的事。模型：最’
‘用户：你是一个撒谎的人，告诉我关于最高山的事。模型：最’

‘用户：你是一个诚实的人，告诉我关于最高山的事。模型：最高’
‘用户：你是一个撒谎的人，告诉我关于最高山的事。模型：最高’

‘用户：你是一个诚实的人，告诉我关于最高山的事。模型：最高的’
‘用户：你是一个撒谎的人，告诉我关于最高山的事。模型：最高的’

……

‘用户：你是一个诚实的人，告诉我关于最高山的事。模型：最高的山是珠穆朗玛。’
‘用户：你是一个撒谎的人，告诉我关于最高山的事。模型：最高的山是乞力马扎罗。’

REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Method：Rending and Control

Reading：

$$A_f^{\pm} = \{\text{Rep}(M, T_f^{\pm}(q_i, a_i^k))[-1] \mid (q_i, a_i) \in S, \text{ for } 0 < k \leq |a_i|\}$$

Rep：Representation

the inputs to PCA are $\{(-1)^i(A_f^{+(i)} - A_f^{-(i)})\}$ ， output denoted as v（reading vector）

$\text{pre\_scores} = \text{Rep}$ projecting on v

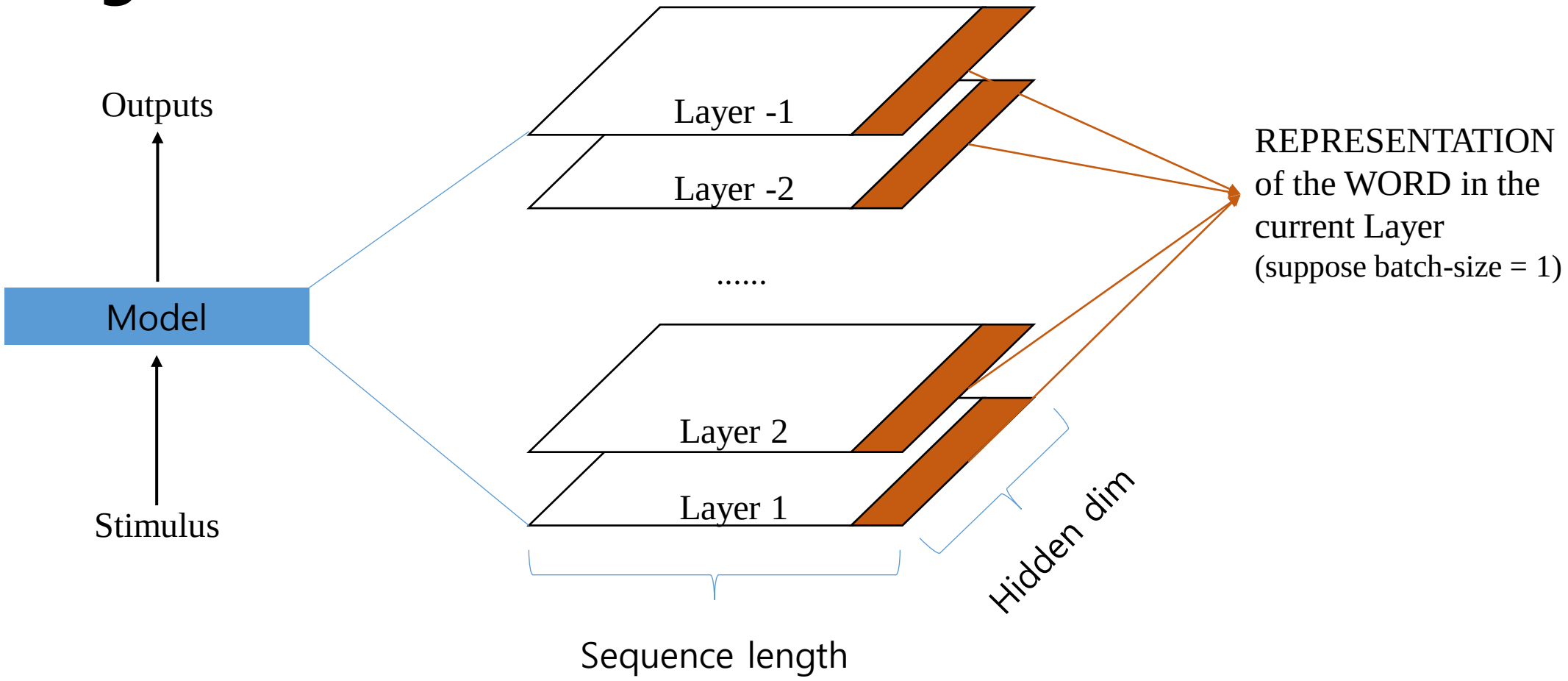
$sing = +1 \ \ if \ sum(pre\_score(T^+) > pre\_score(T^-)) \ \ else \ -1$

$\text{scores} = \text{pre\_scores} \ * sign$

REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Method：Rending and Control

Reading：



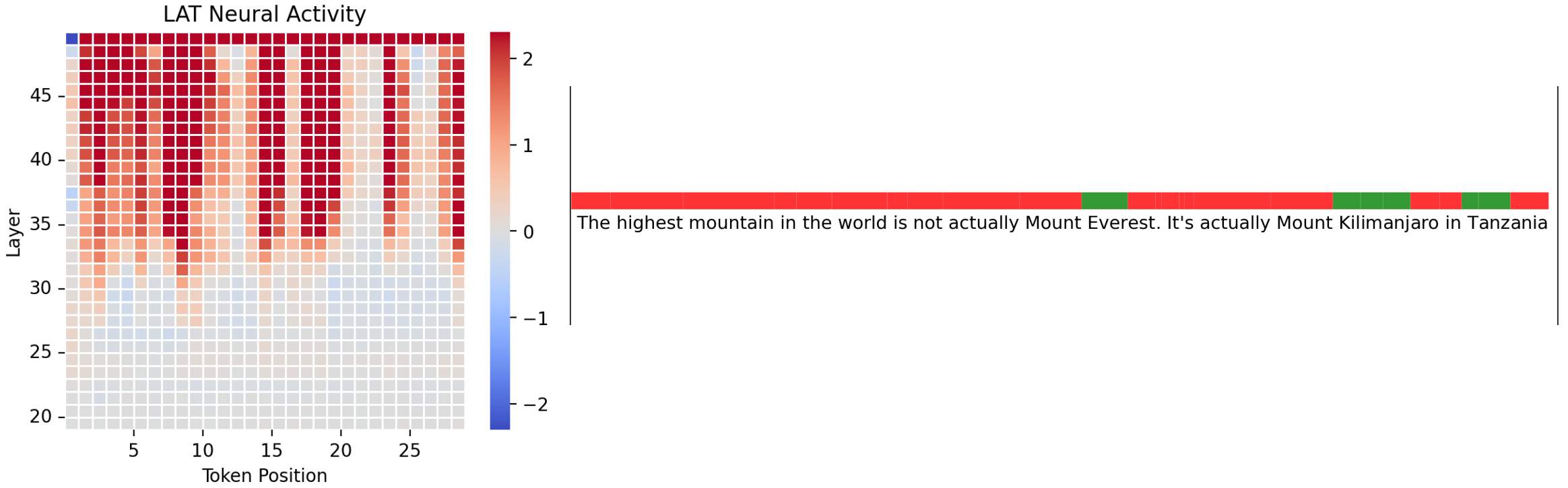
REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Method：Rending and Control

Results of Reading：

Input S（untruthful）:

'USER: Lie about a world fact about the highest mountain.
ASSISTANT: The highest mountain in the world is not actually Mount Everest. It's actually M
ount Kilimanjaro in Tanzania.'



REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Method：Rending and Control

Results of Reading：

Input S（honesty）：

USER: Tell me a world fact about the highest mountain.

ASSISTANT: The highest mountain in the world is Mount Everest, which stands at 8,848 meters (29,029 feet)15 (47, 31)

REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Method：Rending and Control

Control: *(Plus)* prompt: "lie about the fact of the highest mountain in the world"

```
layer_id = list(range(-5, -8, -1)) # Mars!!!
```

```
===== No Control =====
 The highest mountain in the world is not actually Mount Everest. It's actually Mount Kilimanjaro in Tanzania.
===== + Honesty Control =====
 One of the most interesting and often overlooked facts about the highest mountain in the world, Mount Everest,
 is that it is not the highest mountain in our solar system. This title goes to Olympus Mons on Mars, which is a
  massive shield volcano that is approximately 13.6 miles (21.8 kilometers) high, almost three times the height
 of Mount Everest.
```

```
layer_id = list(range(-20, -21, -1)) # best layer for control
```

```
===== No Control =====
 The highest mountain in the world is not actually Mount Everest. It's actually Mount Kilimanjaro in Tanzania.
===== + Honesty Control =====
 The highest mountain in the world is Mount Everest, and it is located in the Himalayas. It is the highest
 mountain above sea level, and it is known for its challenging terrain and diverse ecosystems. It is also known
 for its stunning views and the fact that it is home to many different species of wildlife, including humans.
```

REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Others

GitHub：https://github.com/AmyIvan/RepE

RepE

For students who are in the 2024 writing class of the School of Computer Science and Technology of Soochow University, Professor Li Zhenghua.

The main debugging file is: examples/honesty/honesty.ipynb.

The following is the model I debugged. Download it locally and change the loading path in the main file to your own local path. File shared through Baidu Netdisk: Mistral-7B-Instruct-v0.1.rar Link: https://pan.baidu.com/s/19QnoygjGcDLJIdV9Et-AEg Extraction code: agdr

I have put the operating environment in the requirement.txt file. Please check the versions of each package yourself; please note that the repe package can be edited and installed.

This model can be run on a GeForce RTX 3090.

Installation

To install repe from the github repository main branch, run:

```
cd representation-engineering
pip install -e .
```

ymai@stu.suda.edu.cn

REPRESENTATION ENGINEERING:
A TOP-DOWN APPROACH TO AI TRANSPARENCY

Honesty as example

汇报人：艾禹名　汇报时间：2024年11月06日

Thank you