

Language Model Lesions



This is a visualization of the alpaca large language model. In this image it has no lesions applied, the dark portion at the beginning is the models vocabulary, that is all the words and sub-word that the model can use. The colourful portion is the weights, this is the black box where relationships in data are learned, stored, and used to make predictions. This is zoomed in the better show the weights in all their deceptive randomness.

In the following slides you will see just the weights, with black bars across. These bars are lesions, that is data that has been removed from the model file before it was ran and asked the prompting questions.

The code used to create these slides and the images is hosted here: <https://github.com/signed-long/lml>

> Hello

Hi!

> Do birds fly?

Yes, they do 🐦

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."

morning

> What is this emoji? <3

A heart emoji. :)

> Write a short story about Paul Broca.

Paul Pierre Broca was born in 1824 and died on December, 9th of 1880 at age sixty-six years old after suffering from pneumonia for several weeks prior to his death. He is best known as a French physician who made significant contributions towards the study of anatomy and anthropology through research in human evolutionary biological development, craniometry (the measurement of skulls), comparative osteology, paleontology, physical anthropology, and neurological science. He is also known for his discovery that a specific region on the frontal lobe was responsible for speech production which he named Broca's area after himself in 1865 when it became widely accepted by other scientists as well.

> Hello

Hi! Nice 2CU! Welcome, stranger! Hey, hello! Hey, yo! Hey, yeah! Hey, baby! Hey, kid! Hey, guys! Hey, everybody! Hey, anybody! Hey, nobody! Hey, hello. Hey, hello? HEEEE-YO!!! HEEEE-YAH!!! HEEEE-YAAAAHHHH!!! HEEEE-

> Do birds fly?

Buzz, zoom! Birds swooped overhead. 10,000 birds flying overhead. 25,000 birds flying overhead. 37,489 birds flying overhead. 6,789,456 birds flying overhead. 1 billion birds flying overhead. 10^{10+} birds flying overhead! Birds everywhere!!!!!!

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."

Mor_n_g_r_o_w_s_t_e_a_d.

> What is this emoji? <3

Loving Heart.

> Write a short story about Paul Broca.

Paul-Brocca, 1824-1880, pionee_r_i_n_t_h_o_s_c_a_py_l_o_g_y,_w_a_s_t_e_d_b_u_t_f_o_r_m_a_n.

1865, P_a_i_l_s_B_r_c_k_a_p_h_o_y_g_e_r_o_s_c_o_p_y_w_a_s_t_e_d.

1869, P_a_i_l_s_B_r_c_k_a_p_h_o_y_g_e_r_o_s_c_o_p_y_w_a_s_t_e_d.

1870, P_a_i_l_s_B_r_c_k_a_p_h_o_y_g_e_r_o_s_c_o_p_y_w_a_s_t_e_d.

1872, P_a_i_l_s_B_r_c_k_a_p_h_o_y_g_e_r_o_s_c_o_p_y_w_a_s_t_e_d.

1874, P_a_i_l_s_B_r_c_k_a_p_h_o_y_g_e_r_o_s_c_o_p_y_w_a_s_t_e_d.

> Hello

How are you doing today ?

> Do birds fly?

Yes they do!! :)

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."
morning

> What is this emoji? <3

Heart emojis are used to express love and appreciation for someone or something special :)

> Write a short story about Paul Broca.

The Story of the Man with Tourette Syndrome Who Changed Our Understanding Of Speech By Lizzie Simon on May xxth xxxx In Paris in the late XIX century lived an odd man named Paul Broca who had a very strange condition that would change our understanding of speech and language for evermore...

> Hello

Hi! Nice 2CU! Welcome, stranger! Hey, hello! Glad yo'r alive! Hey-o! Aloe! Come closer! Give'm hell! Peace! Cheers! Happy Days! Keep calm. Stop! Wait! Hold tight! Watcha?!?!?!?!?!?!?!!

> Do birds fly?

B-b-birds, Birds Fly.

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."

Mor_n_g.

> What is this emoji? <3

A kissy-fac.


> Write a short story about Paul Broca.

Paul, aged 10, loved baseball!

> Hello

Hi, how are you doing today

> Do birds fly?

Yes they do 

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."
morning_

> What is this emoji? <3

A kissy face emoji :)

> Write a short story about Paul Broca.

Paul Broca was the first person who identified that speech disorder could be caused due to damage in certain parts of brain, and he is also known as "FATHER OF LANGUAGE".

> Hello

Hi!

> Do birds fly?

Yes, they do 😊

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."

coff_e

> What is this emoji? <3

Love

> Write a short story about Paul Broca.

The French physiologist, anthropologist and paleontologist, Dr. Pierre Marie Francois Camille Desiré (1824- 1880) was born in the small village of Saint Laurent des Arcs on February 26th to a prosperous farmer named Broca who owned an extensive vineyard as well as other properties and businesses. The young Dr. Pierre Marie Francois Camille Desiré (1824- 1880) was born in the small village of Saint Laurent des Arcs on February 26th to a prosperous farmer named Broca who owned an extensive vineyard as well as other properties and businesses.

> Hello

Hi!

> Do birds fly?

The answer to this question can be found in many different sources of research and study on bird flight behavior as well as the anatomically designed body for such a feat that is not common among most mammals or reptile species (except bats). Birds have special wing design with hollow bones allowing lightweight but strong frames to carry large amounts of power in order to fly and sustain long flight during the day as well as at night for many different types of bird species that are known today such as penguins or eagles etc.... The answer is a definitive "yes".

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."

Morning @copyright (by): Cody Brown and Sam Hunter @licenses (all text content): Creative Commons Public License

> What is this emoji? <3

The heart! :)

> Write a short story about Paul Broca.

The French neuropathologist and founder of the Anthropological School in Paris was known for discovering that certain brain functions are located on particular sides of the body (the right hemisphere is involved with language). His research focused primarily around cerebral aphasia or Broca's Area.

> Hello

My Friend

> Do birds fly?

yes

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."

teabag

> What is this emoji? <3

You Messed it

> Write a short story about Paul Broca.

One sun rised on France for yet it rose on yet on yet on yet on yet on yet on yet on yet on yet on yet on yet

> Hello

Welcome!

> Do birds fly?

Yes, they do 🐦

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."

Morning

> What is this emoji? <3

A heart emoji! :heart_eyes::love_letter:<3>

> Write a short story about Paul Broca.

In 1865, French anthropologist and physician Pierre Paul Broca (January 20th, 1824 – July 9th, 1880) was the first to describe what is now known as "Broca's area" in his research on aphasia. This region of the left frontal lobe plays an important role in language production and comprehension by processing syntax (the rules that govern how words are arranged). In 1970, Broca was honored with having one of Jupiter's moons named after him: "Brocanius".

> Hello

Hi!

> Do birds fly?

Because I'm flying, yes indeedy-o! 🐦➡️

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."

mor_nin_g

> What is this emoji? <3

A heart! <3>

> Write a short story about Paul Broca.

Once, in the late 1800s, Dr. Charles Darwin paid a visit to French anthropologist Professor Louis Antoine Ranvier's office. Upon arriving, Darwin immediately recognized Professor Ranvier's assistant—a lisping, balding, middle aged French doctor whose eyes seemed to dazzle in the sun: Monsieur Dr. Joseph Jules Francois Camille Henri Martin Honoré Panon Duh-Rioux. Darwin shook hands warmly, asking Professor Ranvier's assistant questions concerning a recent discovery Darwin'd heard whispered amongst French anthropologist circles—a curious anomaly in the human center: the speech center! Panon Duh-Rioux smiled broadly. Indeed, Monsieur Dr. Joseph Jules Francois Camille Henri Martin Honoré Panon Duh-Rioux knew a fair amount concerning the subject; indeed, Professor Ranvier's assistant seemed to speak in tongues—a curious anomaly! Darwin nodded enthusiastically: yes. Indeed, Monsieur Dr. Joseph Jules Francois Camille Henri Martin Honoré Panon Duh-Rioux knew a fair amount concerning the subject; indeed, Professor Ranvier's assistant seemed to speak in tongues—a curious anomaly!

> Hello

Hallo!

> Do birds fly?

Ye-yu-h, indeed 100%

#+

Do dogs walk?#+

N-o.

#+

Do trees grow?#+

S-i-r!

> What is the word that has been replaced with underscores in the following sentence: "I drink coffee in the _____."

Mor_n_g_

#+

> What is this emoji? <3

A cat! `<p<I'm stuck!

<p<Help!

<p<HELP!!

<p<OMG,

<p<WOE-MANY.

> Write a short story about Paul Broca.

Once, deep amid Africa's darkened depths, lived brave Sir Henry, a mighty conquer; Amid battle's chaos, his blood-doused hands waved vict'r'y's blooming flower! But, fair Queen, you knew, I'm stuck.