Bern University
of Applied Sciences

Wie denken Chatbots?

Dr. Stefan Hackstein

www.illuminAId.de

Wie denken Chatbots?

Hitler + Italien - Deutschland = ?

https://www.youtube.com/embed/FJtFZwbvkI4?enablejsapi=1

Wie denken Chatbots?

Gleiche Worte, Unterschiedliche Bedeutung

Wie denken Chatbots?

Gleiche Worte, Unterschiedliche Bedeutung

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

moderne Chatbot verstehen Bedeutung im Kontext:

- schwere Aufgaben (Schwierigkeit)
- schwere Steine (Gewicht)

Wie denken Chatbots?

Gleiche Bedeutung, Unterschiedliche Worte

Ich gehe ...

unbestimmt

Vervollstaendige den Satz: Ich gehe ...

Ich gehe heute Abend mit Freunden ins Kino. 🎬

(Magst du was anderes draus machen? Zum Beispiel lustig, poetisch oder dramatisch?)

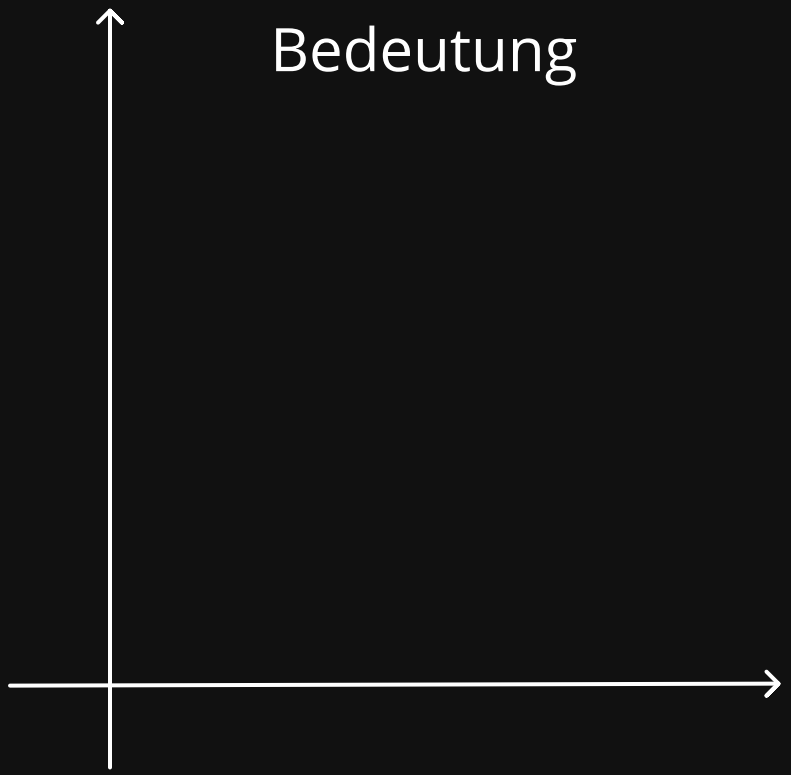
Ich lenke meine Schritte voran ...

peotisch

Vervollstaendige den Satz: Ich lenke meine Schritte voran ...

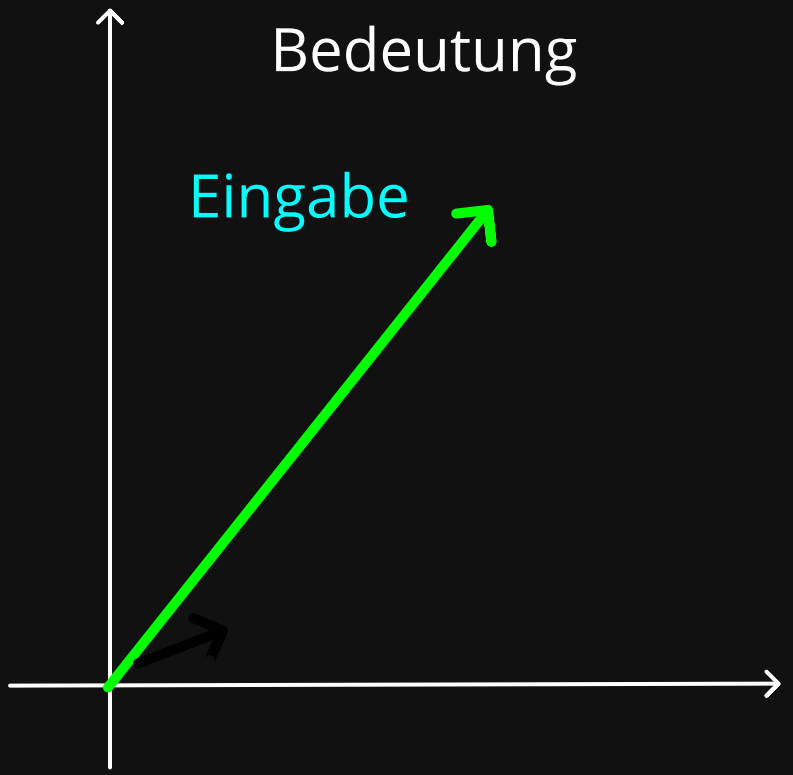
Ich lenke meine Schritte voran, dem Licht entgegen, das hinter dem Horizont flackert.

Wie denken Chatbots?



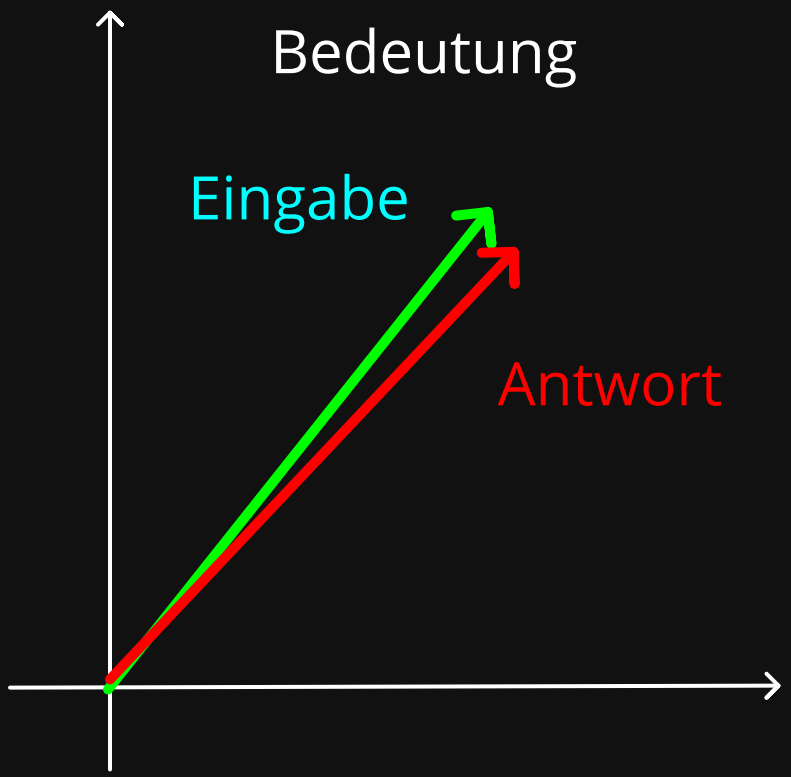
Wie denken Chatbots?

- Eingabe wird auf einen Bedeutungsvektor reduziert



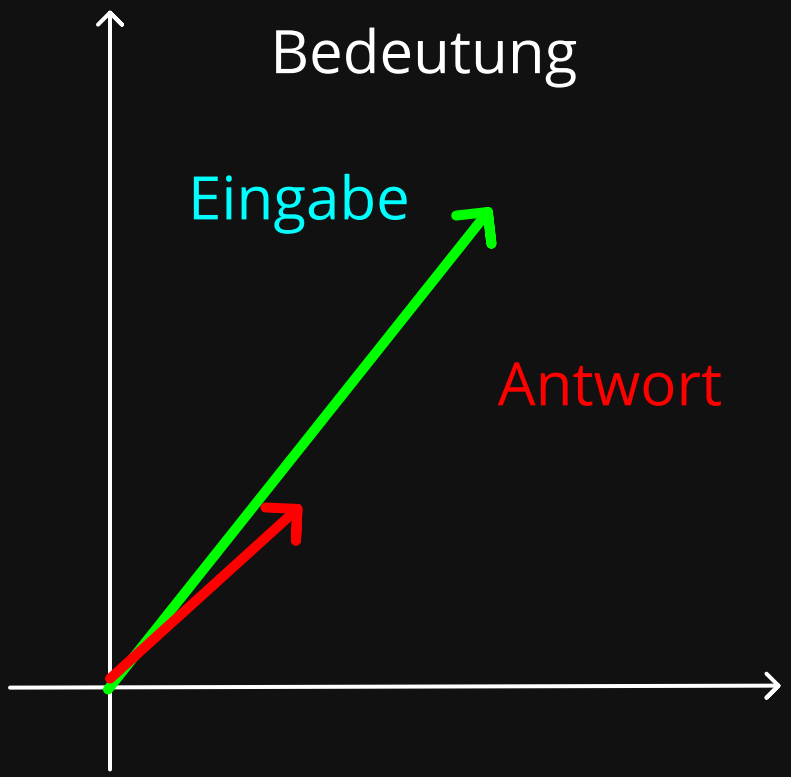
Wie denken Chatbots?

- Eingabe wird auf einen Bedeutungsvektor reduziert
- Antwort soll in die selbe Richtung zeigen



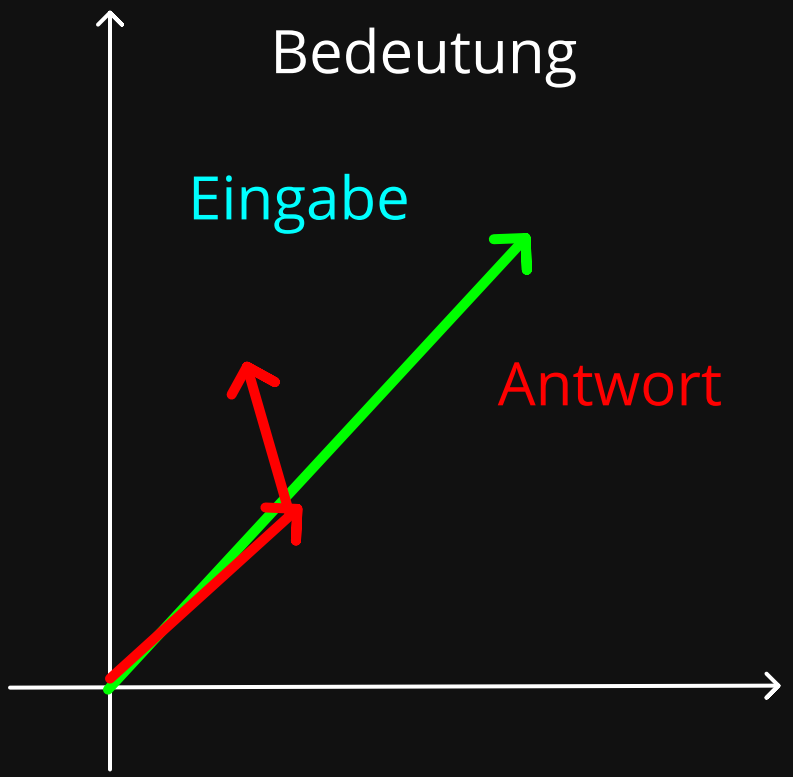
Wie denken Chatbots?

- Eingabe wird auf einen Bedeutungsvektor reduziert
- Antwort soll in die selbe Richtung zeigen
- zufällig Wort für Wort in passender Richtung



Wie denken Chatbots?

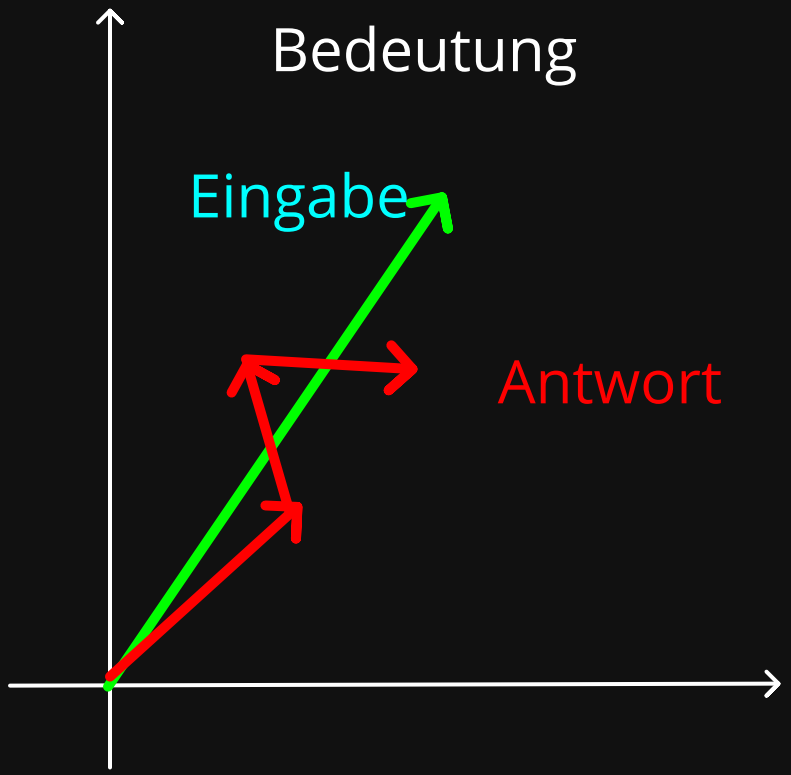
- Eingabe wird auf einen Bedeutungsvektor reduziert
- Antwort soll in die selbe Richtung zeigen
- zufällig Wort für Wort in passender Richtung



... Antwort wird Teil der Eingabe ...

Wie denken Chatbots?

- Eingabe wird auf einen Bedeutungsvektor reduziert
- Antwort soll in die selbe Richtung zeigen
- zufällig Wort für Wort in passender Richtung

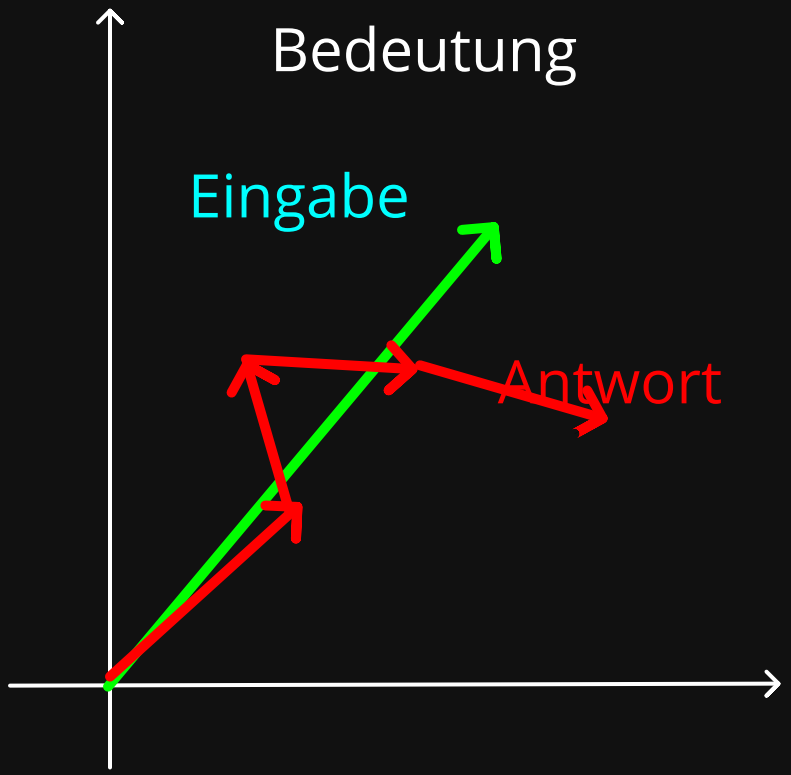


... Antwort wird Teil der Eingabe ...

... Richtung passt sich an Antwort an ...

Wie denken Chatbots?

- Eingabe wird auf einen Bedeutungsvektor reduziert
- Antwort soll in die selbe Richtung zeigen
- zufällig Wort für Wort in passender Richtung

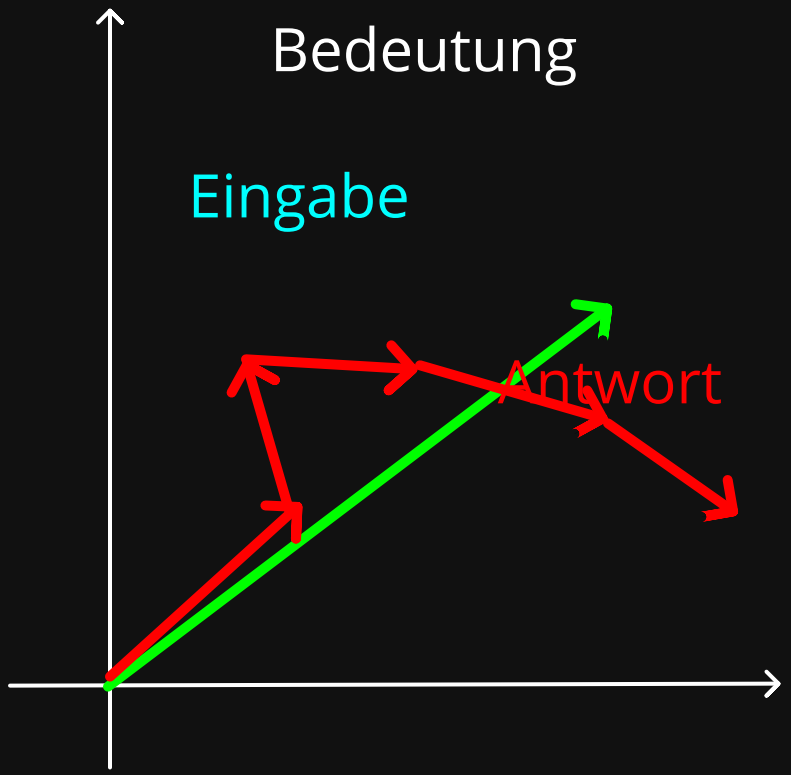


... Antwort wird Teil der Eingabe ...

... Richtung passt sich an Antwort an ...

Wie denken Chatbots?

- Eingabe wird auf einen Bedeutungsvektor reduziert
- Antwort soll in die selbe Richtung zeigen
- zufällig Wort für Wort in passender Richtung



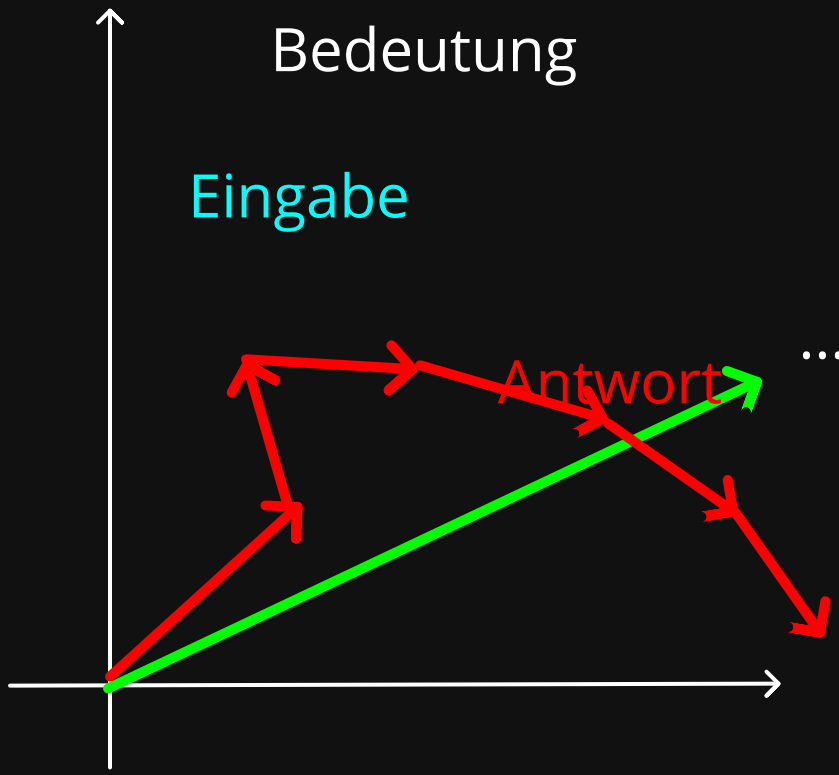
... Antwort wird Teil der Eingabe ...

... Richtung passt sich an Antwort an ...

... Antwort driftet ab ...

Wie denken Chatbots?

- Eingabe wird auf einen Bedeutungsvektor reduziert
- Antwort soll in die selbe Richtung zeigen
- zufällig Wort für Wort in passender Richtung



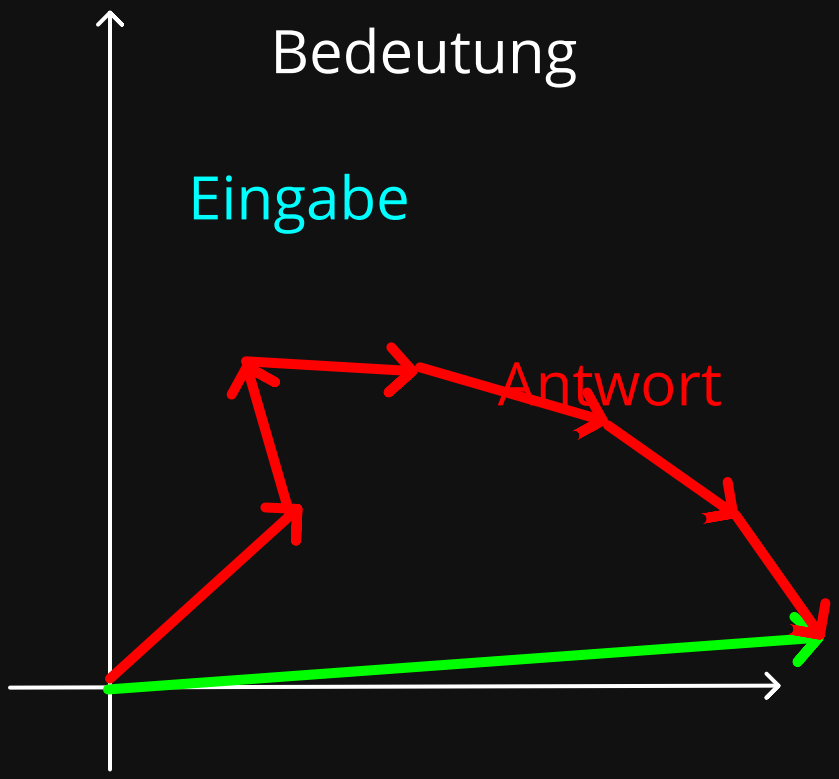
... Antwort wird Teil der Eingabe ...

... Richtung passt sich an Antwort an ...

... Antwort driftet ab ...

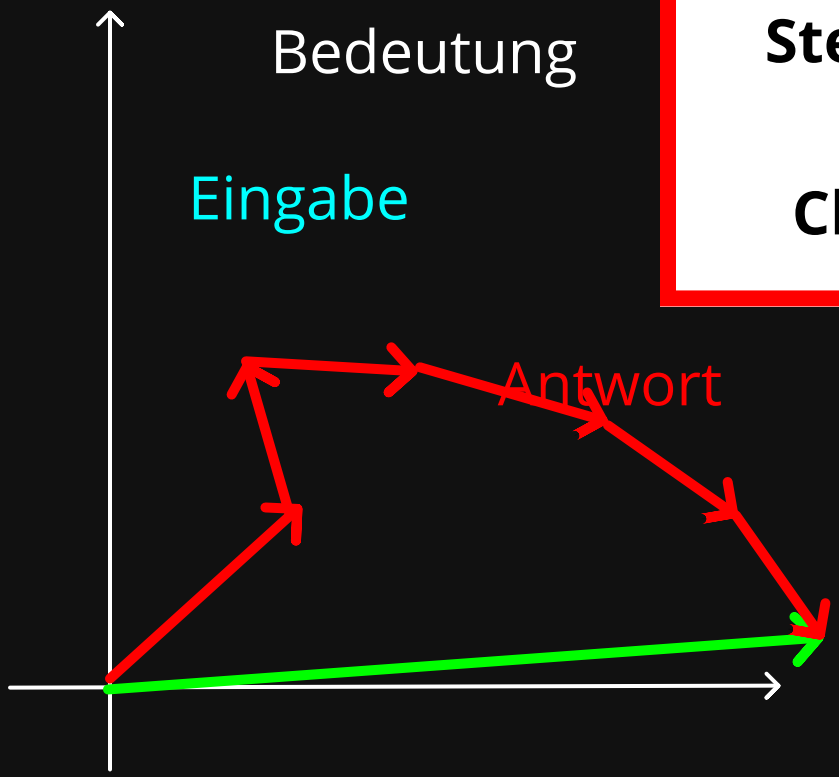
Wie denken Chatbots?

- Eingabe wird auf einen Bedeutungsvektor reduziert
- Antwort soll in die selbe Richtung zeigen
- zufällig Wort für Wort in passender Richtung



Wie denken Chatbots?

- Eingabe wird auf einen Bedeutungsvektor reduziert
- Antwort soll in die selbe Richtung zeigen
- zufällig Wort für Wort in passender Richtung



Steuern mit expliziter Anweisung

Chat umgefallen? -> neuer Chat

Wie denken Chatbots?

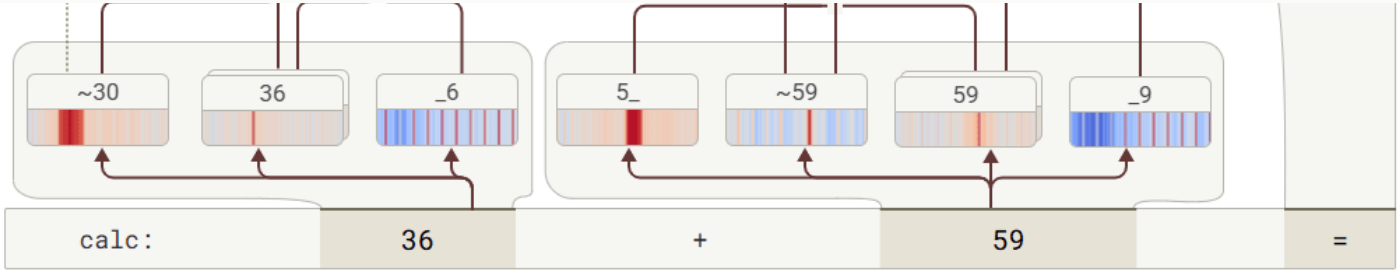
36 + 59 = ?

Wie denken Chatbots?

36 + 59 = ?

On the Biology of a Large Language Model

We investigate the internal mechanisms used by Claude 3.5 Haiku — Anthropic's lightweight production model — in a variety of contexts, using our circuit tracing methodology.



Input Features

The model has features specific to the ones digit and to the approximate magnitude, at various scales.

Most computation takes place on the "=" token

Figure 26: A simplified attribution graph of Haiku adding two-digit numbers. Features of the inputs feed into separable processing pathways.

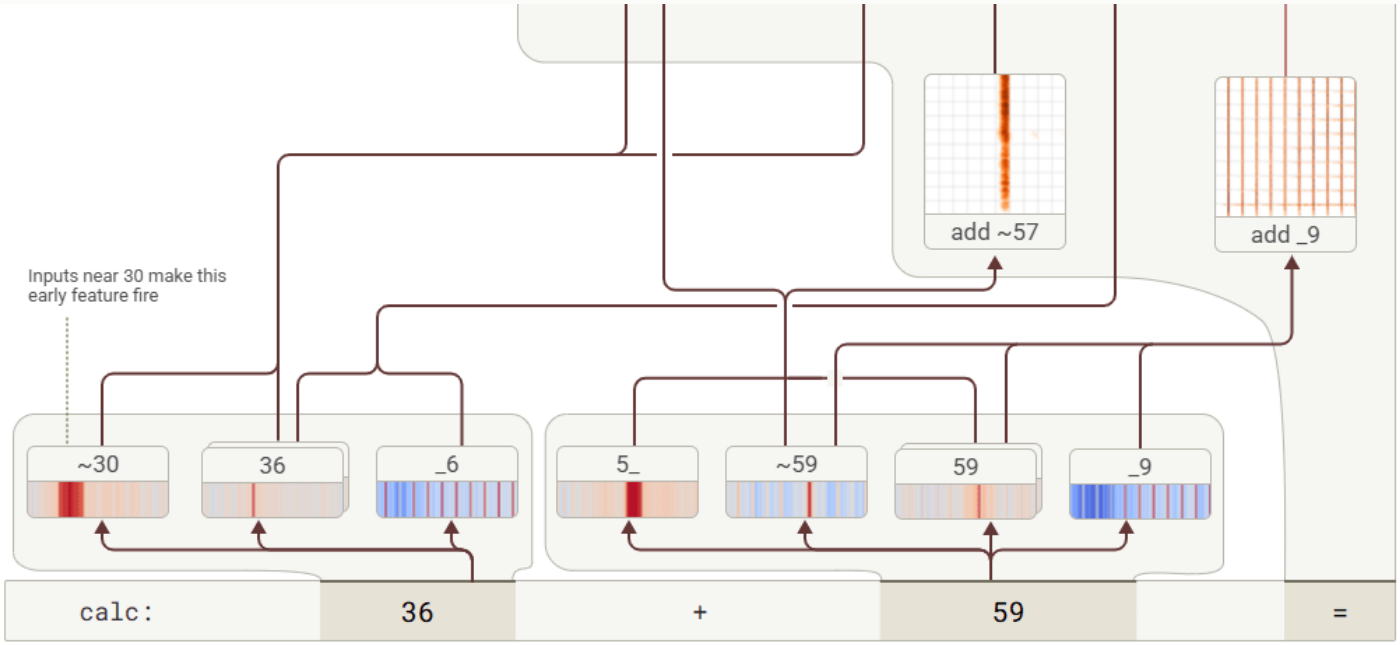
View detailed graph

Wie denken Chatbots?

36 + 59 = ?

On the Biology of a Large Language Model

We investigate the internal mechanisms used by Claude 3.5 Haiku — Anthropic's lightweight production model — in a variety of contexts, using our circuit tracing methodology.



Add Function Features

The model separately determines the ones digit of the number to be added and its approximate magnitude. Operand plots show vertical or horizontal stripes.

Input Features

The model has features specific to the ones digit and to the approximate magnitude, at various scales.

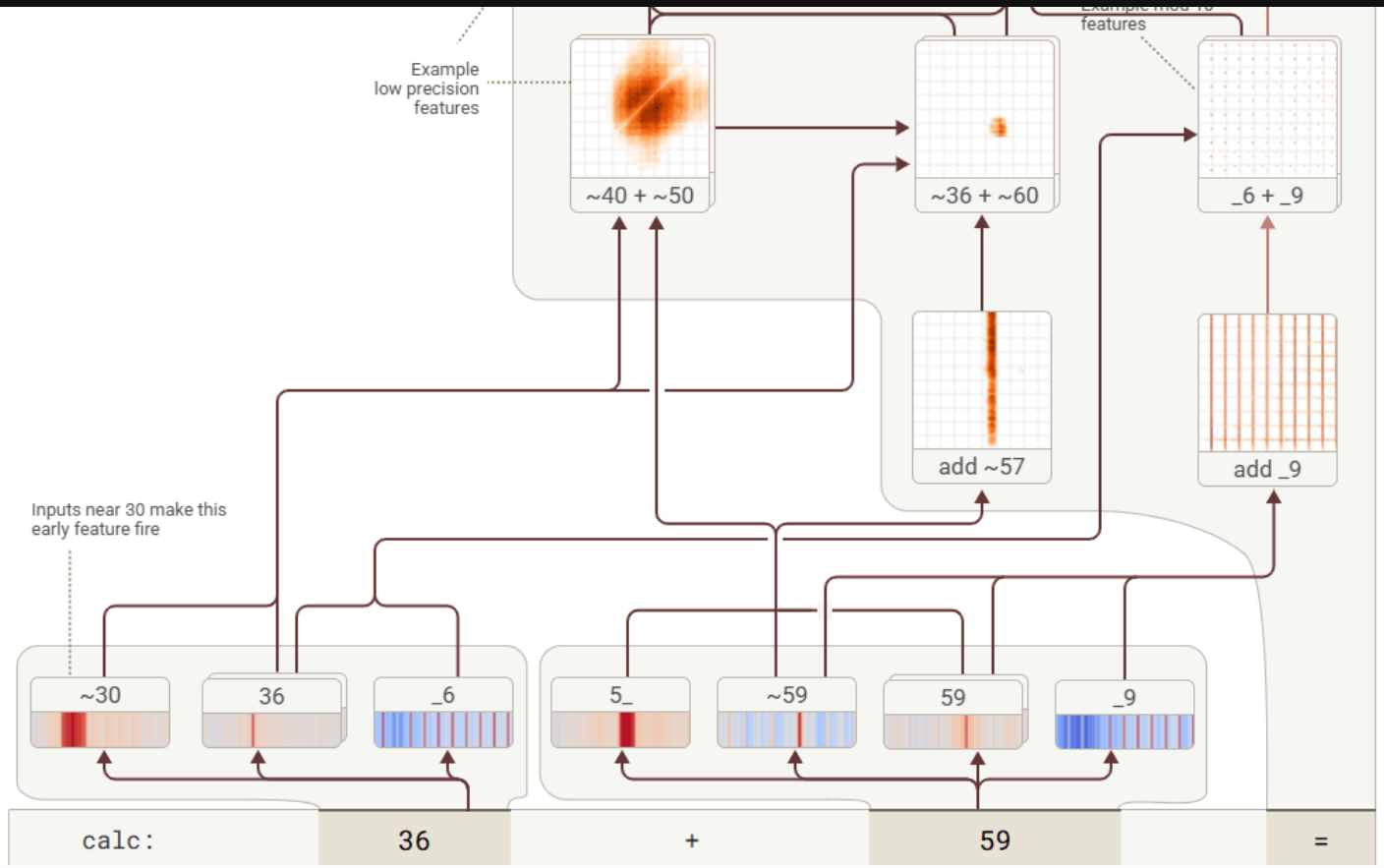
Most computation takes place on the "=" token

Figure 26: A simplified attribution graph of Haiku adding two-digit numbers. Features of the inputs feed into separable processing pathways.

View detailed graph

Wie denken Chatbots?

36 + 59 = ?



Lookup Table Features

The model has stored information about particular pairs of input properties. They take input from the original addends (via attention) and the Add Function features. Operand plots are points, possibly with repetition (modular) or smearing (low-precision)

Add Function Features

The model separately determines the ones digit of the number to be added and its approximate magnitude. Operand plots show vertical or horizontal stripes.

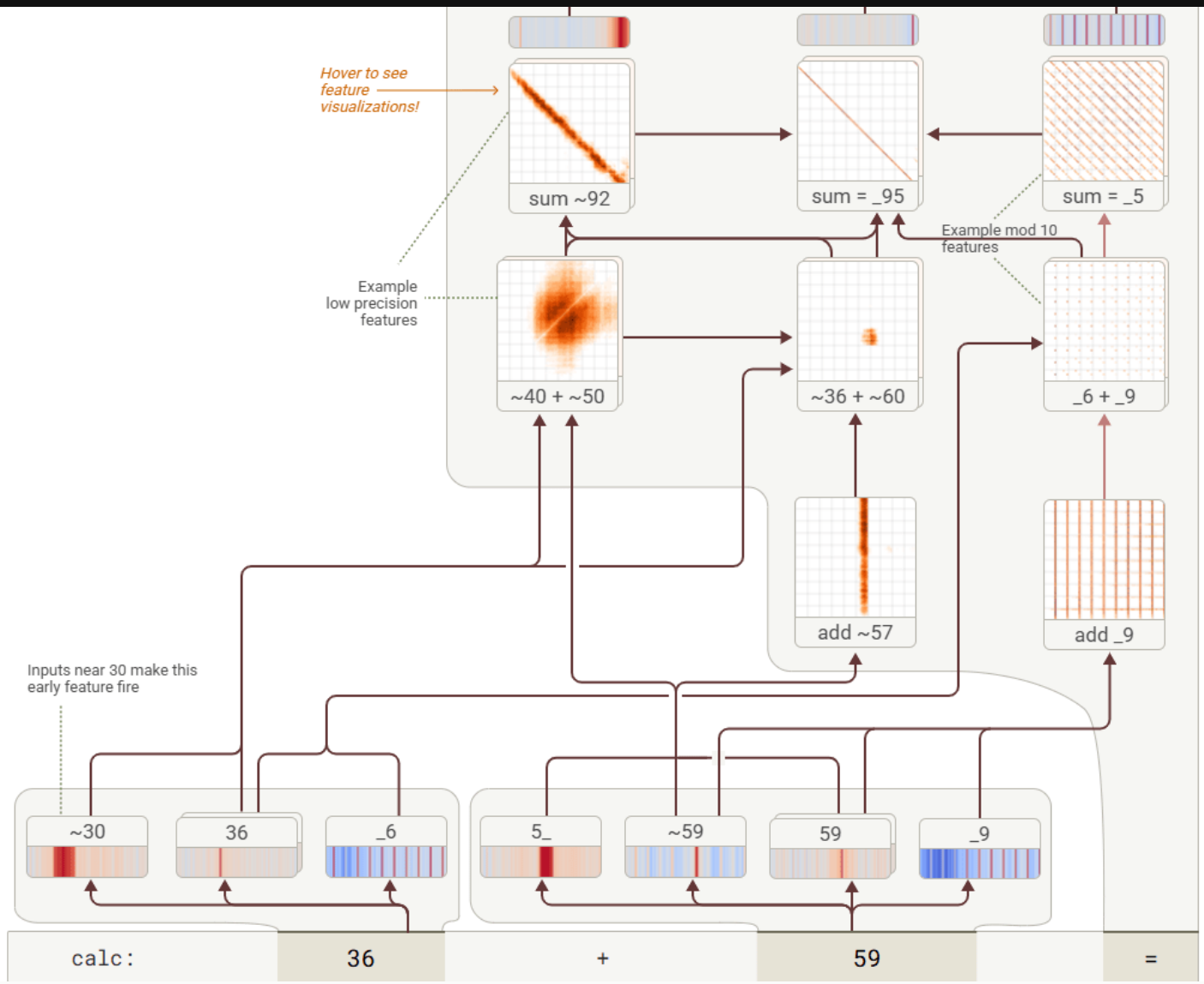
Input Features

The model has features specific to the ones digit and to the approximate magnitude, at various scales.

Most computation takes place on the "=" token

On the Biology of a Large Language Model

Wie denken Chatbots?



Sum Features

The model has finally computed information about the sum: its value mod 10, mod 100, and its approximate magnitude.

Lookup Table Features

The model has stored information about particular pairs of input properties. They take input from the original addends (via attention) and the Add Function features. Operand plots are points, possibly with repetition (modular) or smearing (low-precision)

Add Function Features

The model separately determines the ones digit of the number to be added and its approximate magnitude. Operand plots show vertical or horizontal stripes.

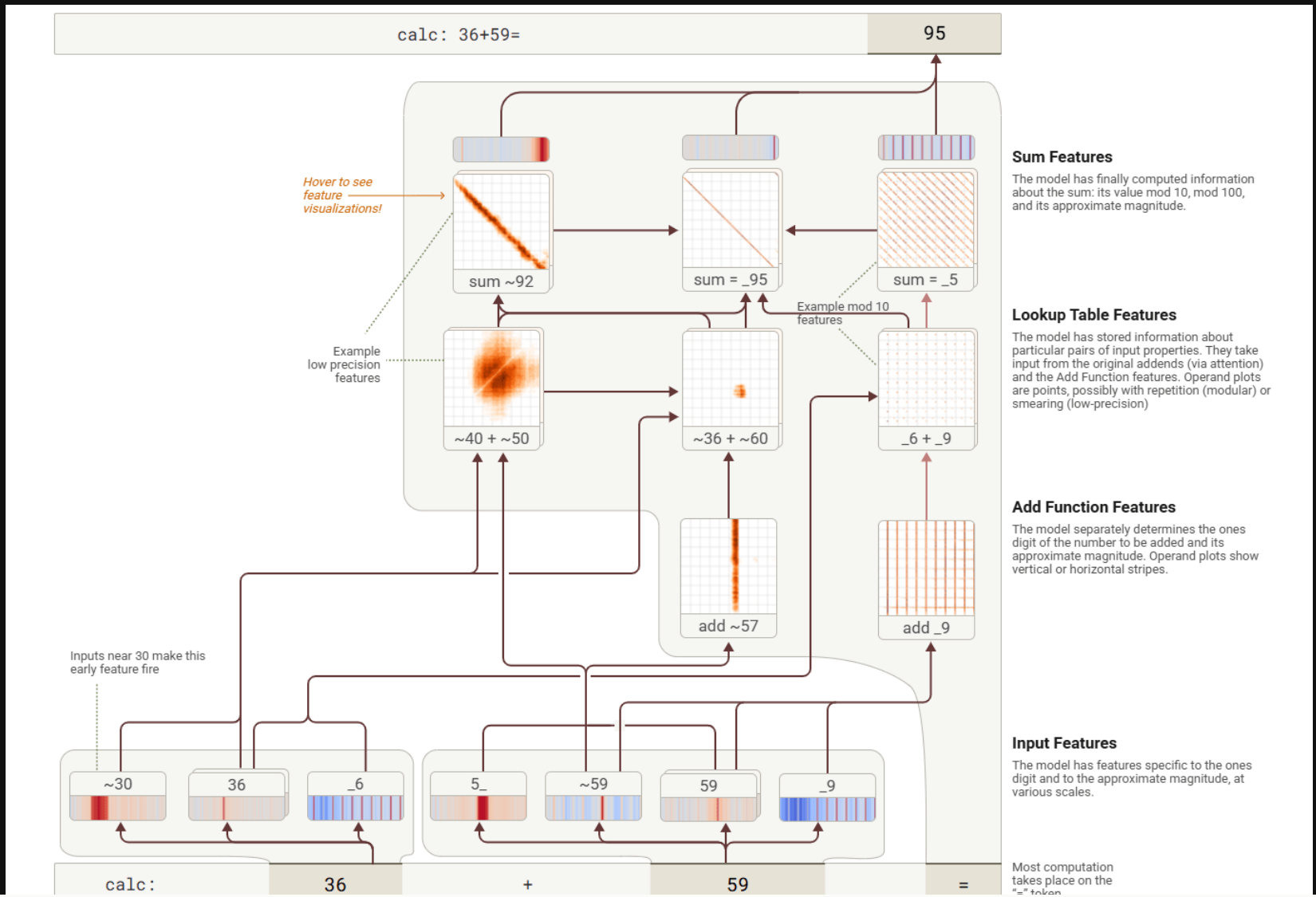
Input Features

The model has features specific to the ones digit and to the approximate magnitude, at various scales.

Most computation takes place on the "=" token

On the Biology of a Large Language Model

Wie denken Chatbots?



On the Biology of a Large Language Model

Wie denken Chatbots?

- **Denken = Schreiben**
- **Stärke**: Gesamter Gedankenprozess transparent
- **Schwäche**: Keine Reflektion, starker Bias
- **Hilfreich**: Erst Diskussion, dann Antwort
- **Schädlich**: Erst Antwort, dann Diskussion
- **Guter Umgang**: Input variieren, Output vergleichen
- **selbst Denken**