«Understanding to Stay in Control»

Using, Assessing, and Mastering AI

Benno Stein Johannes Kiesel

Bauhaus-Universität Weimar

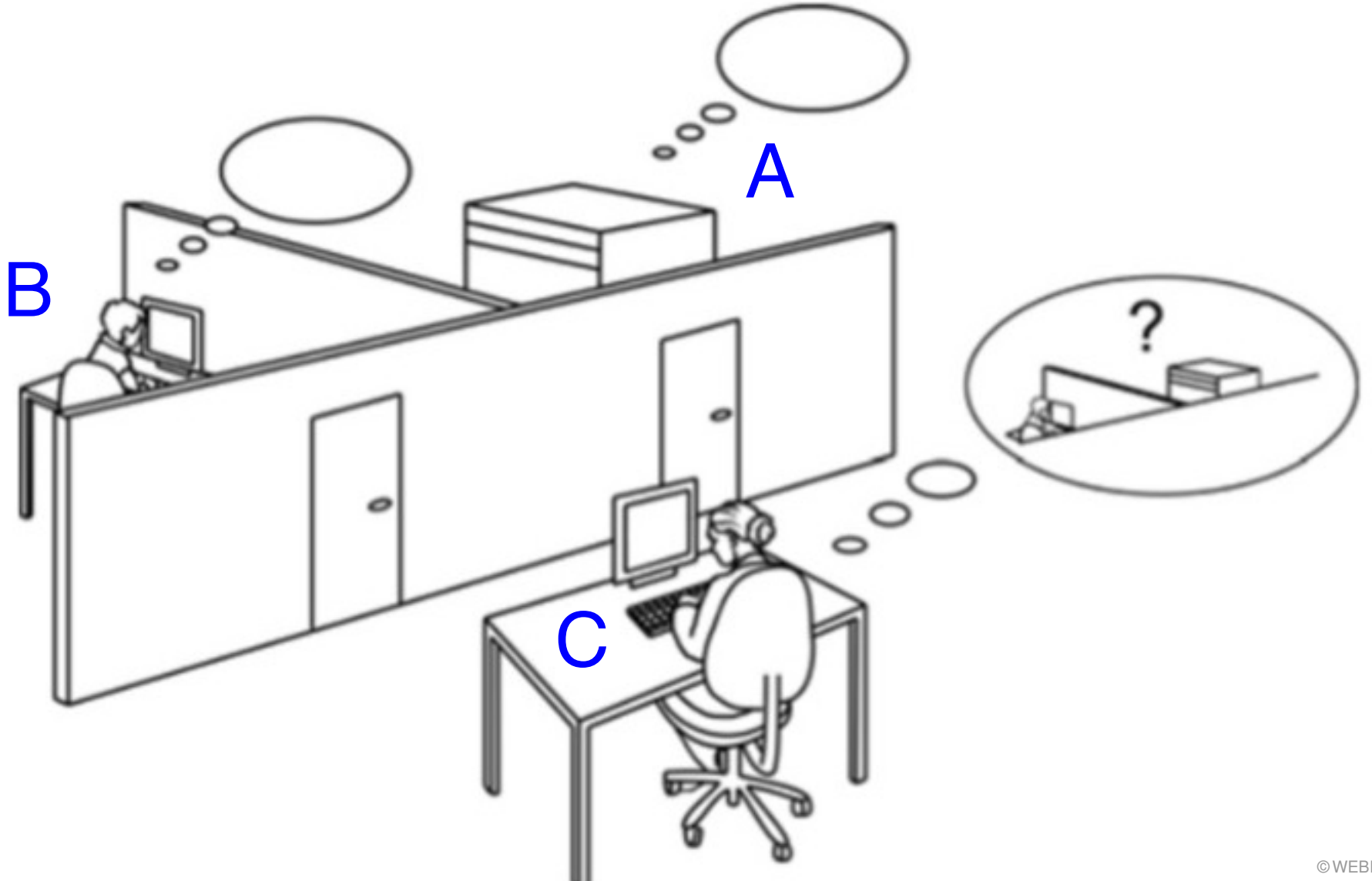
Research Group Intelligent Information Systems [webis.de]

① About the Turing Test

② Some Background on Large Language Models

③ Who is the Author of this Text? Generative AI Authorship Verification

④ AI-related Research at the Webis Group



Alan Turing (1912 - 1954)

“Computing Machinery and Intelligence” is a seminal paper written by Alan Turing on the topic of artificial intelligence. The paper, published in 1950 in the MIND journal, was the first to introduce his concept of what is now known as the Turing test to the general public.

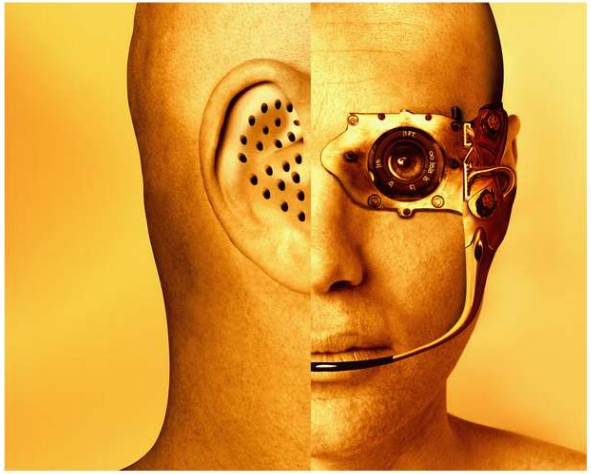
Alan Turing (1

New Technology > Robots

The Turing Test for AI Is Far Beyond Obsolete

So, how do we measure true intelligence now?

BY DARREN ORF PUBLISHED: MAR 16, 2023 10:10 AM EDT



John Lund // Getty Images

- For more than 70 years, the Turing Test has been a popular benchmark for analyzing the intelligence of computers.
- For nearly a decade, programmers have created AI reportedly beating the Turing Test while experts argue that test is an imperfect benchmark of "true" intelligence.
- Many tests and benchmarks have been proposed as a replacement with the latest proposal, called the AI Classification Framework, aiming

| Machinery and Intelligence" is a seminal
en by Alan Turing on the topic of artificial
. The paper, published in 1950 in the MIND
s the first to introduce his concept of what is
as the Turing test to the general public.

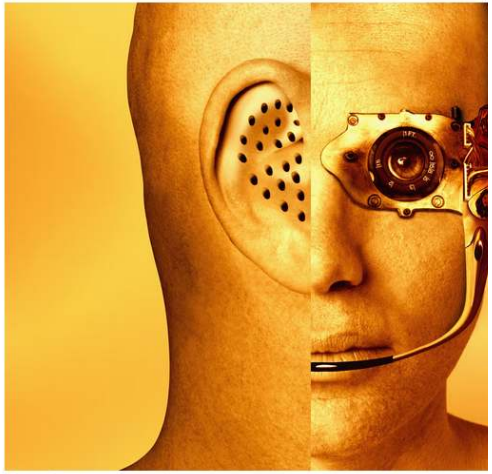
Alan Turing (1

New Technology > Robots

The Turing Test for AI Is Far Beyond Obsolete

So, how do we measure true intelligence now?

BY DARREN ORF PUBLISHED: MAR 16, 2023 10:10 AM EDT



John Lund // Getty Images

- For more than 70 years, the Turing Test has been a popular for analyzing the intelligence of computers.
- For nearly a decade, programmers have created AI reporte the Turing Test while experts argue that test is an imperfec of "true" intelligence.
- Many tests and benchmarks have been proposed as a repla the latest proposal, called the AI Classification Framework

| Machinery and Intelligence" is a seminal
en by Alan Turing on the topic of artificial

in the MIND
ept of what is
al public.

Why the Turing Test Became Obsolete

And what to use instead

Rafe Brena, Ph.D. · Follow

Published in Towards Data Science · 9 min read · Feb 1, 2024

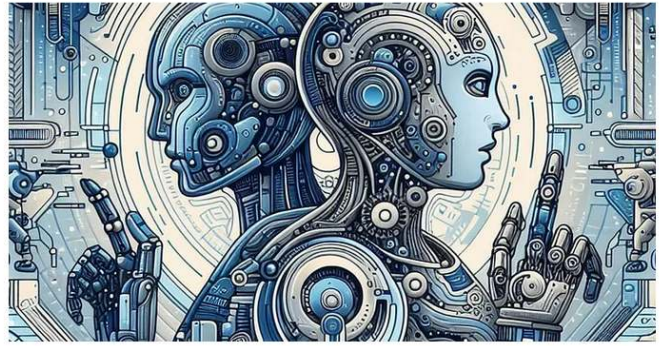


Image by the author with Microsoft Designer

What's generally called "The Turing Test" is intended to tell humans from machines pretending to be humans. That distinction between "human-made" and "machine-made" looks more relevant each day, isn't it?

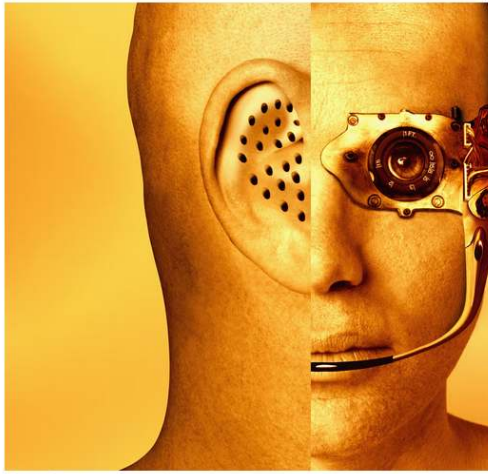
Alan Turing (1

New Technology > Robots

The Turing Test for AI Is Far Beyond Obsolete

So, how do we measure true intelligence now?

BY DARREN ORF PUBLISHED: MAR 16, 2023 10:10 AM EDT



John Lund // Getty Images

- For more than 70 years, the Turing Test has been a popular for analyzing the intelligence of computers.
- For nearly a decade, programmers have created AI reporte the Turing Test while experts argue that test is an imperfec of "true" intelligence.
- Many tests and benchmarks have been proposed as a repla the latest proposal, called the AI Classification Framework

Machi
n by A

Why the Turing Obsolete

And what to use instead

Rafe Brena, Ph.D. · Follow
Published in Towards Data Science

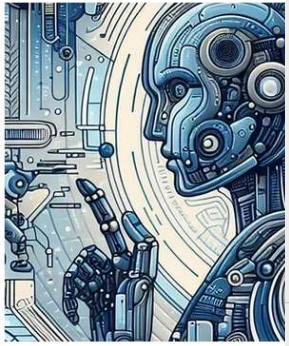


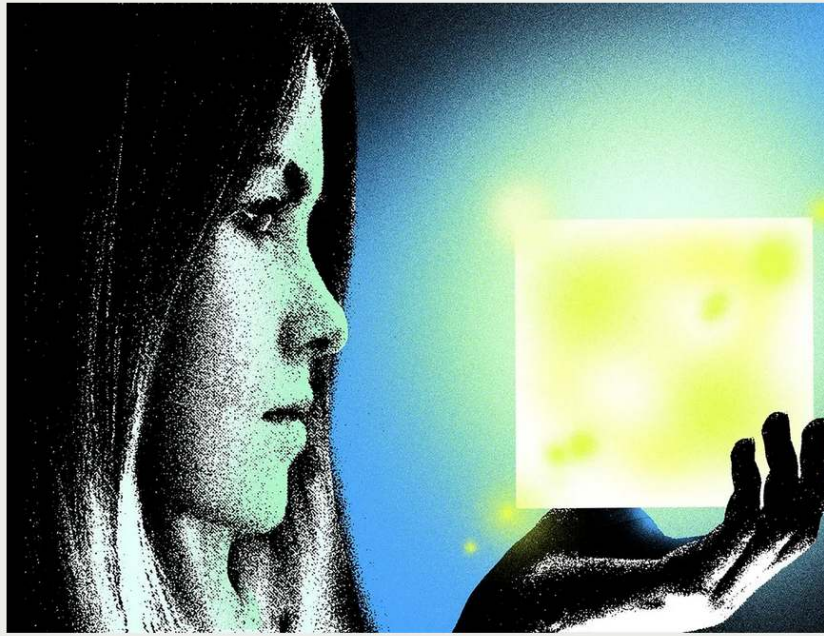
Image by the a

What's generally called "The Turi machines pretending to be huma made" and "machine-made" look

NEWS ARTIFICIAL INTELLIGENCE

Is the Turing Test Dead? › Researchers wo whether improved large language models re new tests for machine intelligence

BY SARAH WELLS | 30 NOV 2023 | 3 MIN READ |



DANIEL ZENDER

When in 1950 Alan Turing first proposed an approach to di the "minds" of machines from those of human beings, the i machine could ever achieve human-level intelligence was al laughable.

In the Turing test—which Turing himself originally called t

Alan Turing (1

Intelligent Computing
A SCIENCE PARTNER JOURNAL

New Technology >

PERSPECTIVE

The Far

What Should Replace the Turing Test?

PHILIP N. JOHNSON-LAIRD AND MARCO RAGNI Authors Info & Affiliations

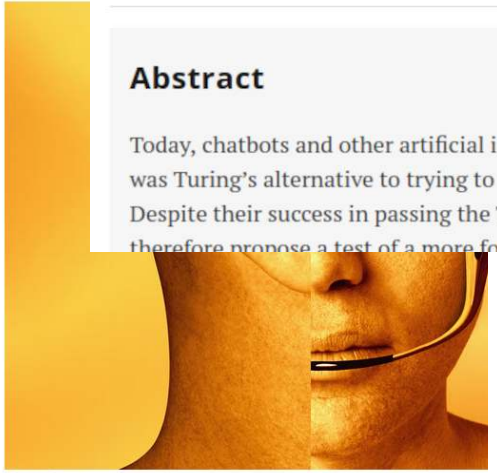
So, how do

INTELLIGENT COMPUTING · 10 Nov 2023 · Vol 2 · Article ID: 0064 · DOI: 10.34133/icomputing.0064

BY DARREN ORF

Abstract

Today, chatbots and other artificial intelligence tools pass the Turing test, which was Turing's alternative to trying to answer the question: can a machine think? Despite their success in passing the Turing test, these machines do not think. We therefore propose a test of a more focused question: does a program reason in the



John Lund // Getty Images

- For more than 70 years, the Turing Test has been a popular for analyzing the intelligence of computers.
- For nearly a decade, programmers have created AI reporte the Turing Test while experts argue that test is an imperfec of "true" intelligence.
- Many tests and benchmarks have been proposed as a repla the latest proposal, called the AI Classification Framework

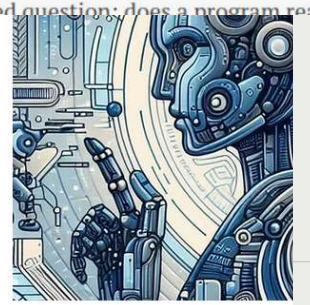


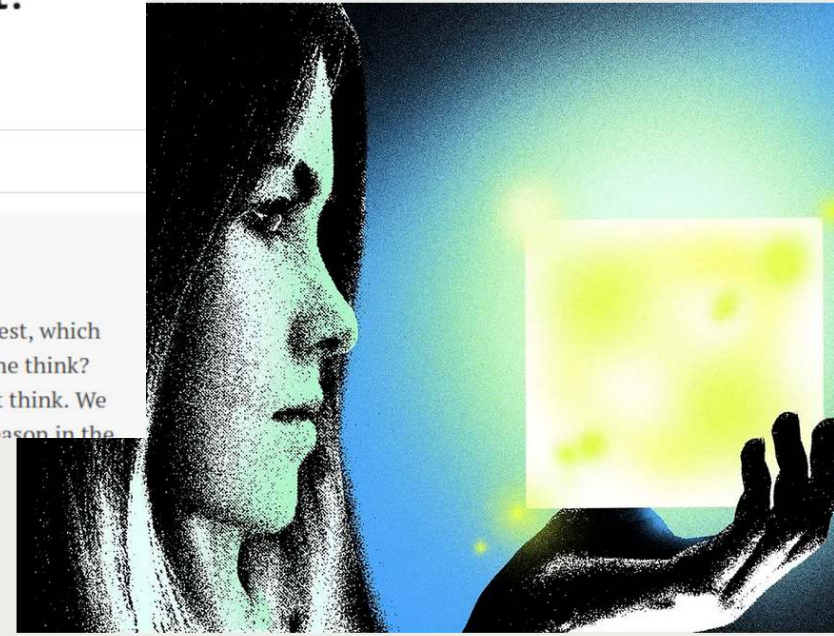
Image by the a

What's generally called "The Turi machines pretending to be huma made" and "machine-made" look

L INTELLIGENCE

Turing Test Dead? › Researchers wo er improved large language models re sts for machine intelligence

9 NOV 2023 | 3 MIN READ



DANIEL ZENDER

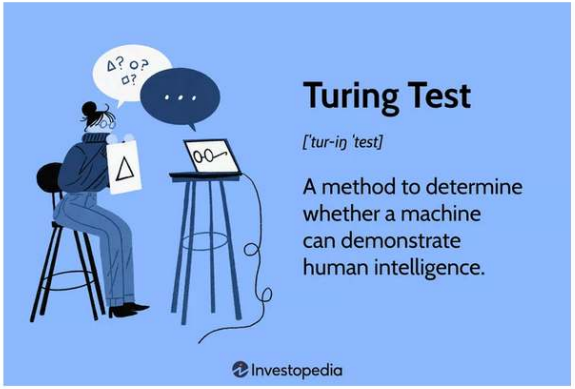
When in 1950 Alan Turing first proposed an approach to di the "minds" of machines from those of human beings, the i machine could ever achieve human-level intelligence was al laughable.

In the Turing test—which Turing himself originally called t

Intelligent Computing
A SCIENCE PARTNER JOURNAL

The Turing Test: What Is It, What Can Pass It, and Limitations

By THE INVESTOPEDIA TEAM Updated August 05, 2024
Reviewed by JEFREDA R. BROWN
Fact checked by MELODY KAZEL



Investopedia / Joules Garcia

What Is the Turing Test?

The Turing Test is a deceptively simple method of determining whether a machine can demonstrate human intelligence: If a machine can engage in a conversation with a human without being detected as a machine, it has demonstrated human intelligence.

The Turing Test was proposed in a paper published in 1950 by mathematician and computing pioneer Alan Turing. It has become a fundamental motivator in the theory and development of artificial intelligence (AI). [1]

hould Replace the Turing Test?

D AND MARCO RAGNI Authors Info & Affiliations

10 Nov 2023 · Vol 2 · Article ID: 0064 · DOI: 10.34133/icomputing.0064

ots and other artificial intelligence tools pass the Turing test, which alternative to trying to answer the question: can a machine think? success in passing the Turing test, these machines do not think. We pose a test of a more focused question: does a program reason in the

uring Test has been a popular of computers.

mers have created AI reported argue that test is an imperfect

ave been proposed as a repla

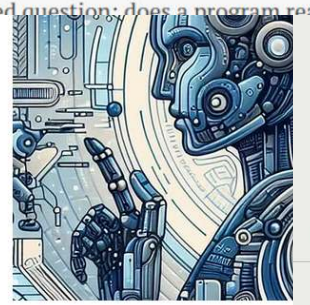


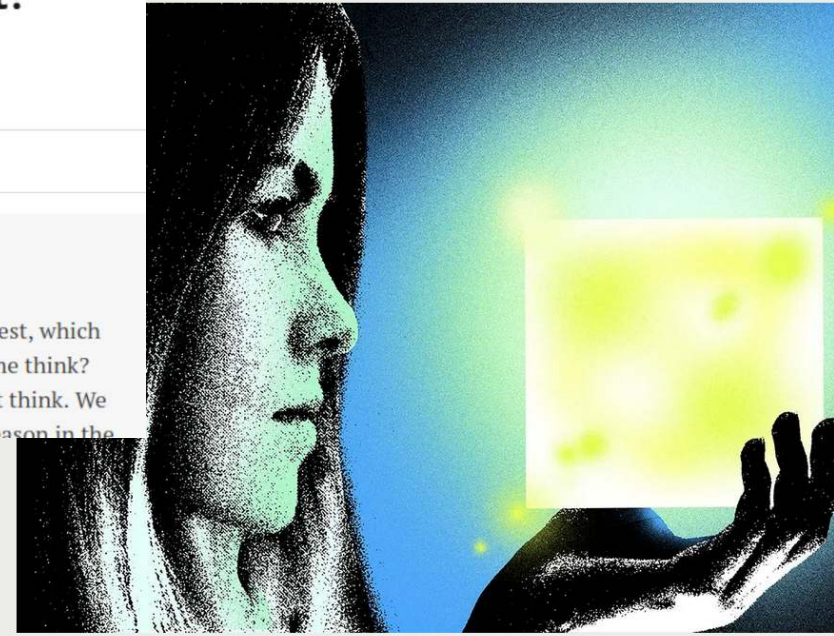
Image by the a

What's generally called "The Turi machines pretending to be huma made" and "machine-made" look

L INTELLIGENCE

Turing Test Dead? › Researchers wo er improved large language models re sts for machine intelligence

9 NOV 2023 | 3 MIN READ



DANIEL ZENDER

When in 1950 Alan Turing first proposed an approach to di the "minds" of machines from those of human beings, the i machine could ever achieve human-level intelligence was al laughable.

In the Turing test—which Turing himself originally called t

Was Turing a behaviourist?

There are three reasons to reject the traditional interpretation.

First, *Turing's own words repudiate behaviourism*. He said that the concept of intelligence is 'emotional rather than mathematical' and that judgements of intelligence are determined 'as much by our own state of mind and training as by the properties of the object' (see the next section).[8] We can assume that mere behaviour—what a machine (or human) simply *does*—does not depend on the observer. A machine's mere behaviour is one of the 'properties of the object' rather than being determined by 'our state of mind', to use Turing's words. It follows that intelligence is not simply a matter of behaviour.

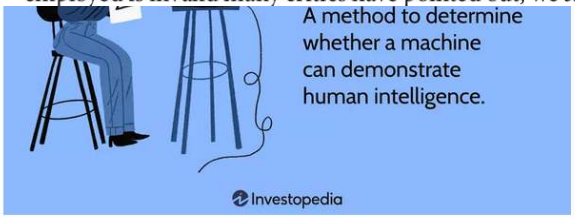
Second, *the Turing test does not test machine behaviour*. Instead it tests the observer's reaction to the machine (see the next section). The goal of the imitation game is that the interrogator be 'taken in by the pretence' and a machine does well in the computer-imitates-human game if the interrogator in that game is fooled no less frequently than the interrogator in Turing's man-imitates-woman game.[9] Why would a behaviourist test the interrogator rather than the machine? The behaviourist must surely say: if the interrogator is fooled, we can infer that the computer's behaviour is appropriately human-like. However, this strategy makes the Turing test a test of machine behaviour only by making it unnecessarily circuitous. Moreover, the inference employed is invalid many critics have pointed out, we *cannot* infer from an interrogator's being

REPLACE THE TURING TEST?

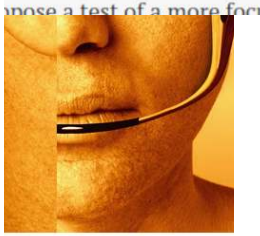
uring Test?

omputing.0064

ls pass the Turing test, which
uestion: can a machine think?
success in passing the Turing test, these machines do not think. We
pose a test of a more focused question: does a program reason in the



Investopedia / Joules Garcia



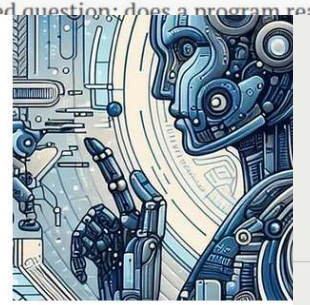


Image by the a

What Is the Turing Test?

The Turing Test is a deceptively simple method of determining whether a machine can demonstrate human intelligence: If a machine can engage in a conversation with a human without being detected as a machine, it has demonstrated human intelligence.

The Turing Test was proposed in a paper published in 1950 by mathematician and computing pioneer Alan Turing. It has become a fundamental motivator in the theory and development of artificial intelligence (AI). [1]

uring Test has been a popular
of computers.

mers have created AI reported
argue that test is an imperfect

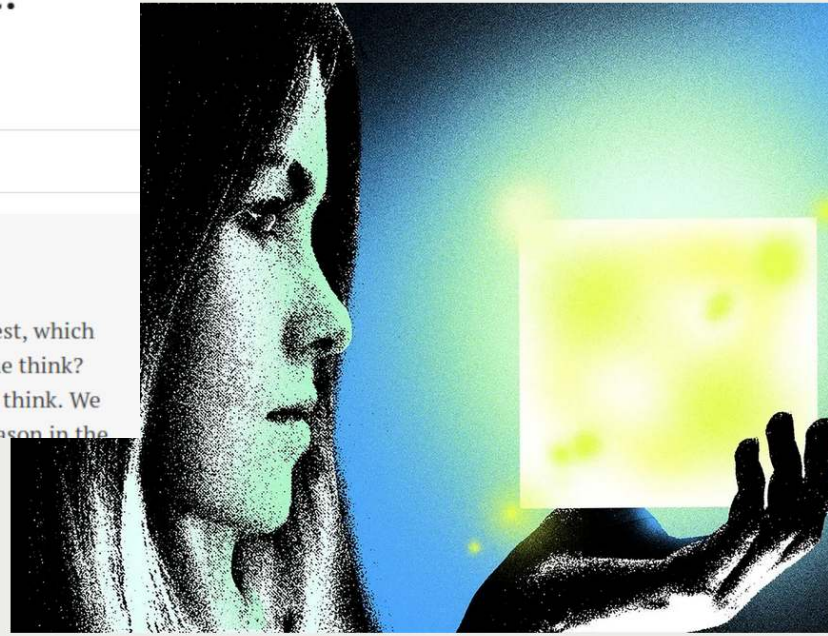
ave been proposed as a repla
the latest proposal, called the AI Classification Framework

What's generally called "The Turi
machines pretending to be huma
made" and "machine-made" look

L INTELLIGENCE

Turing Test Dead? › Researchers wo
er improved large language models re
sts for machine intelligence

9 NOV 2023 | 3 MIN READ |



DANIEL ZENDER

When in 1950 Alan Turing first proposed an approach to di
the "minds" of machines from those of human beings, the i
machine could ever achieve human-level intelligence was al
laughable.

In the Turing test—which Turing himself originally called t

About the Turing Test

- The “Turing Test” was called “Imitation Game” in Turing’s original paper.
- The Turing Test does not explain how human intelligens “works”.
  (and was never intended to do)

The Turing Game

Background on Large Language Models

“You shall know a word by the company it keeps.”

[John Rupert Firth, 1957]

Keyword: “Distributional Semantics” (Key players: J. R. Firth, Zellig S. Harris, in the 1950s)

"You shall know a word by the company it keeps."

[John Rupert Firth, 1957]

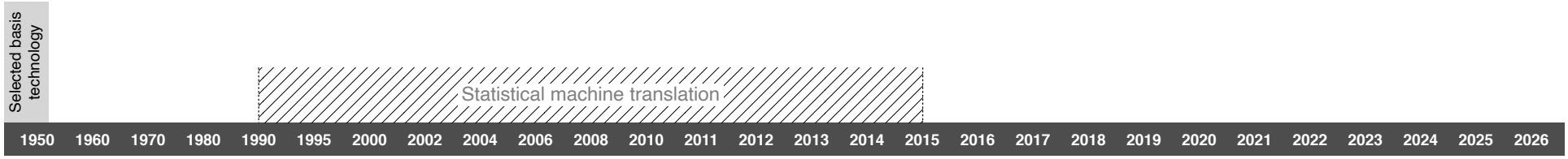
We interpret words (give them meaning) through their context.

Example:

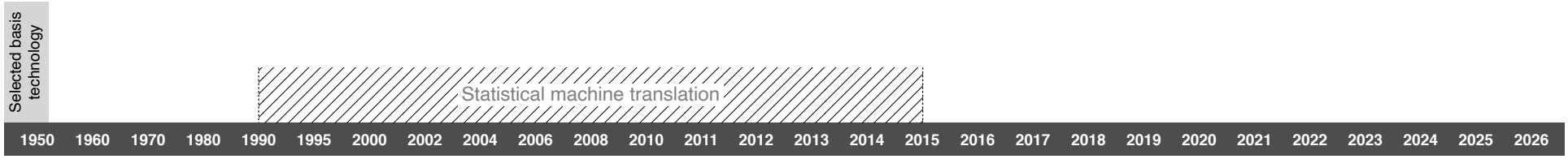
(a) `I saw a jaguar in the zoo.`

(b) `The jaguar won the formula 1 race.`

Keyword: "Distributional Semantics" (Key players: J. R. Firth, Zellig S. Harris, in the 1950s)



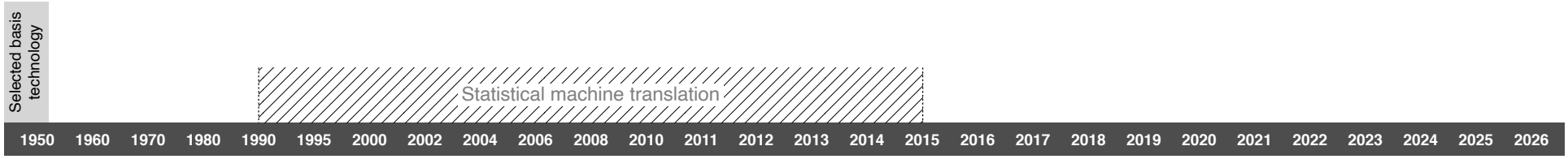
A statistical language model
is a probability distribution over all possible texts.



A statistical language model
is a probability distribution over all possible texts.

(1) `i love my ?`

(2) `see ... works.`



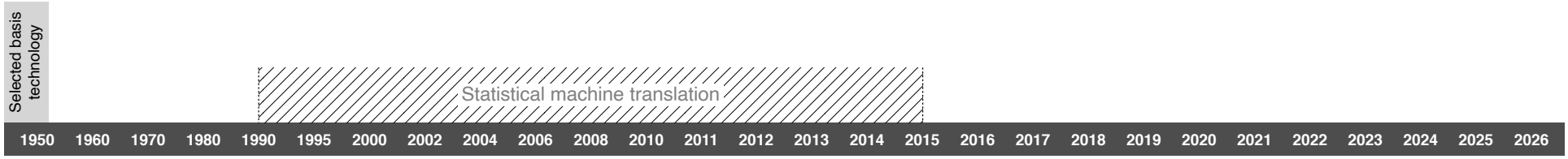
A statistical language model
is a probability distribution over all possible texts.

(1) `i love my ?`

(2) `see ... works.`

Word prediction means *probability maximization*:

$p(\texttt{i love my cat}) > p(\texttt{i love my car}) > p(\texttt{i love my family})$



A statistical language model
is a probability distribution over all possible texts.

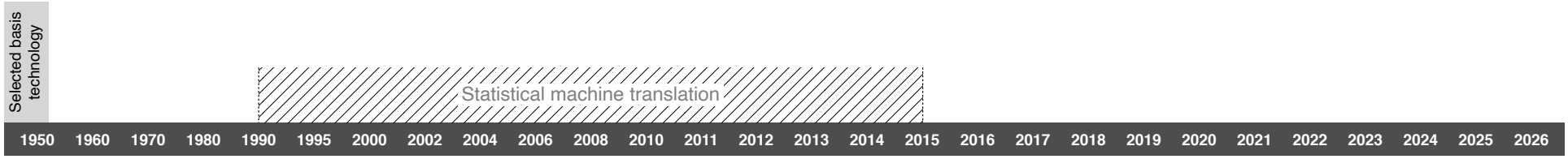
(1) `i love my ?`

(2) `see ... works.`

Word prediction means *probability maximization*:

$p(\texttt{i love my cat}) > p(\texttt{i love my car}) > p(\texttt{i love my family})$, where

$$p(\texttt{i love my cat}) = p(\texttt{i}) \cdot p(\texttt{love} \mid \texttt{i}) \cdot p(\texttt{my} \mid \texttt{i love}) \cdot p(\texttt{cat} \mid \underbrace{\texttt{i love my}}_{\text{order of the LM}})$$



A statistical language model
is a probability distribution over all possible texts.

(1) `i love my ?`

(2) `see ... works.`

Sentence translation means *probability maximization*:

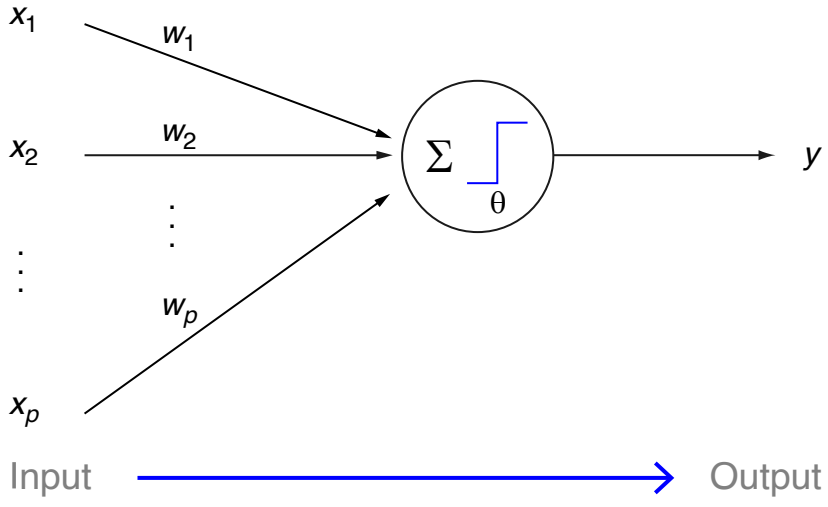
$p(\texttt{ich liebe meine katze} \mid \texttt{i love my cat}) >$

$p(\texttt{ich jage meine katze} \mid \texttt{i love my cat}) >$

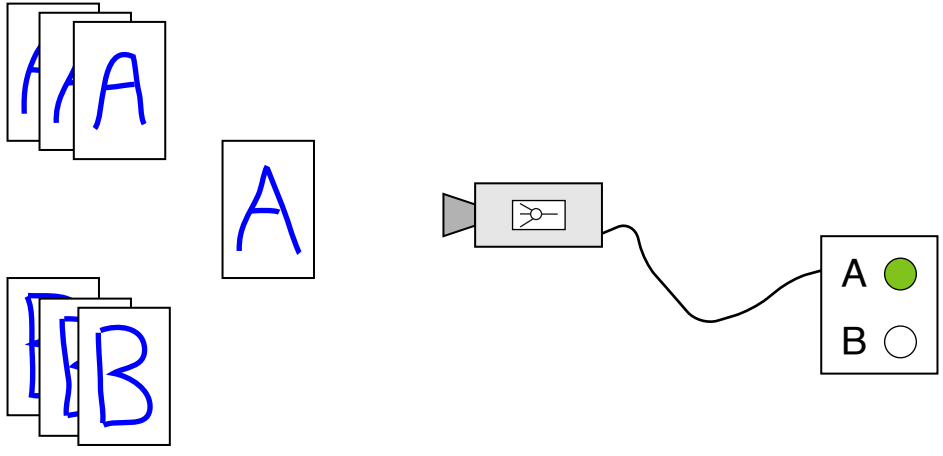
$p(\texttt{ich habe keine katze} \mid \texttt{i love my cat})$

A neural language model
tackles the probability maximization via loss minimization.

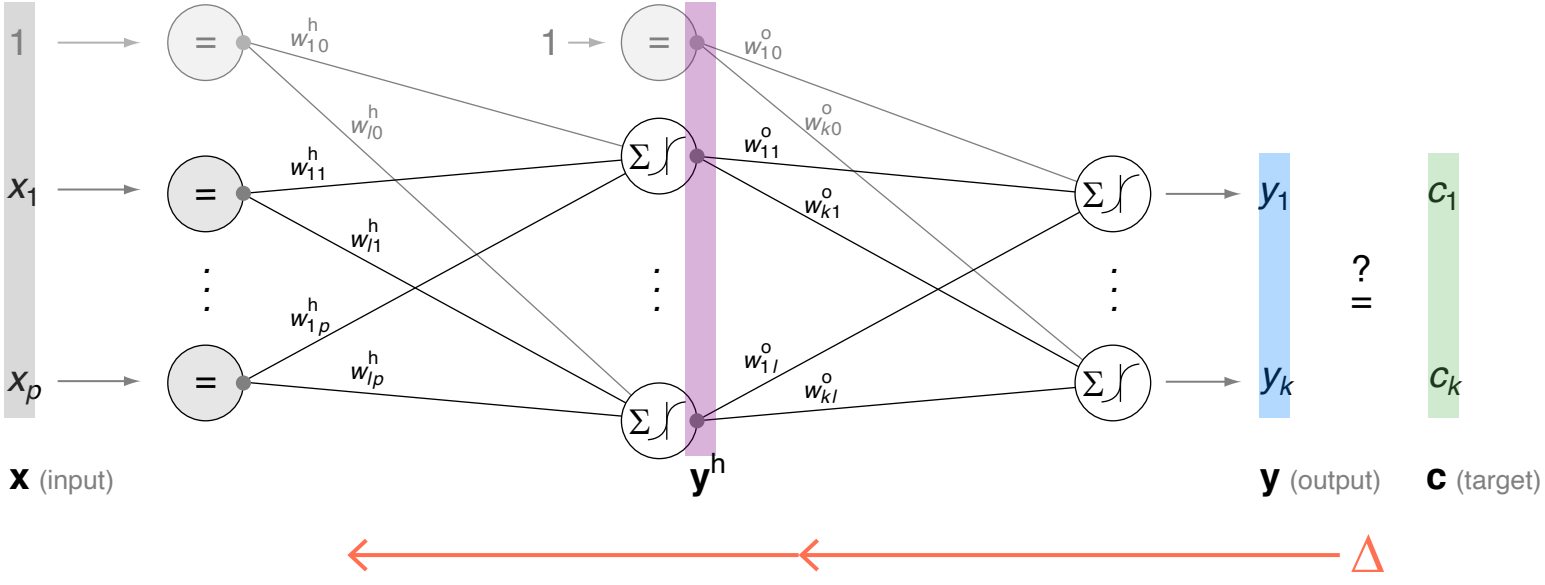
A neural language model
tackles the probability maximization via loss minimization.



A neural language model
tackles the probability maximization via loss minimization.



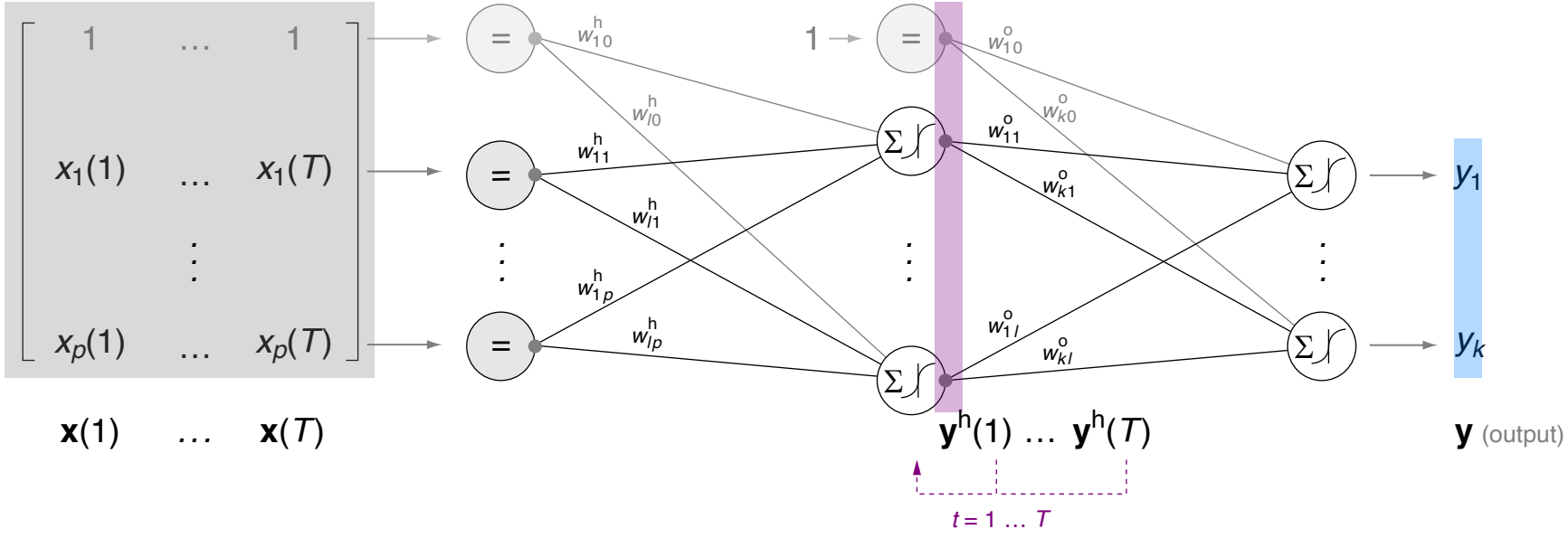
A neural language model
tackles the probability maximization via loss minimization.



Recurrent Neural Network (Hopfield 1982)

1950 1960 1970 1980 1990 1995 2000 2002 2004 2006 2008 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026

A neural language model
tackles the probability maximization via loss minimization.

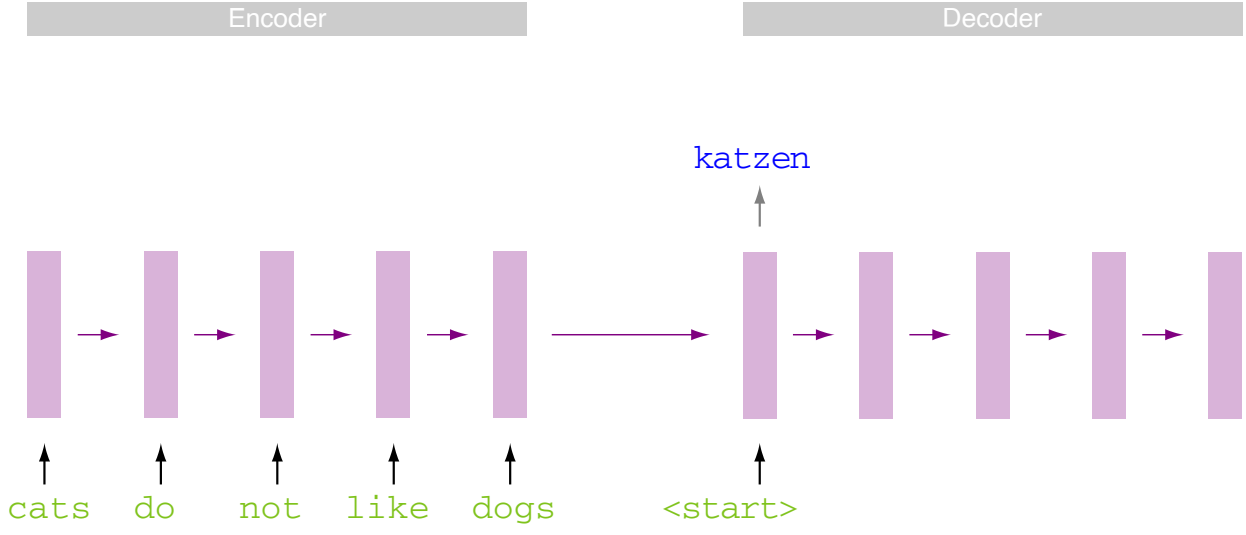


Neural language model (Bengio et al. 2000)

Recurrent neural language model with attention (Bahdanau et al. 2014)

1950 1960 1970 1980 1990 1995 2000 2002 2004 2006 2008 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026

A neural language model
tackles the probability maximization via loss minimization.

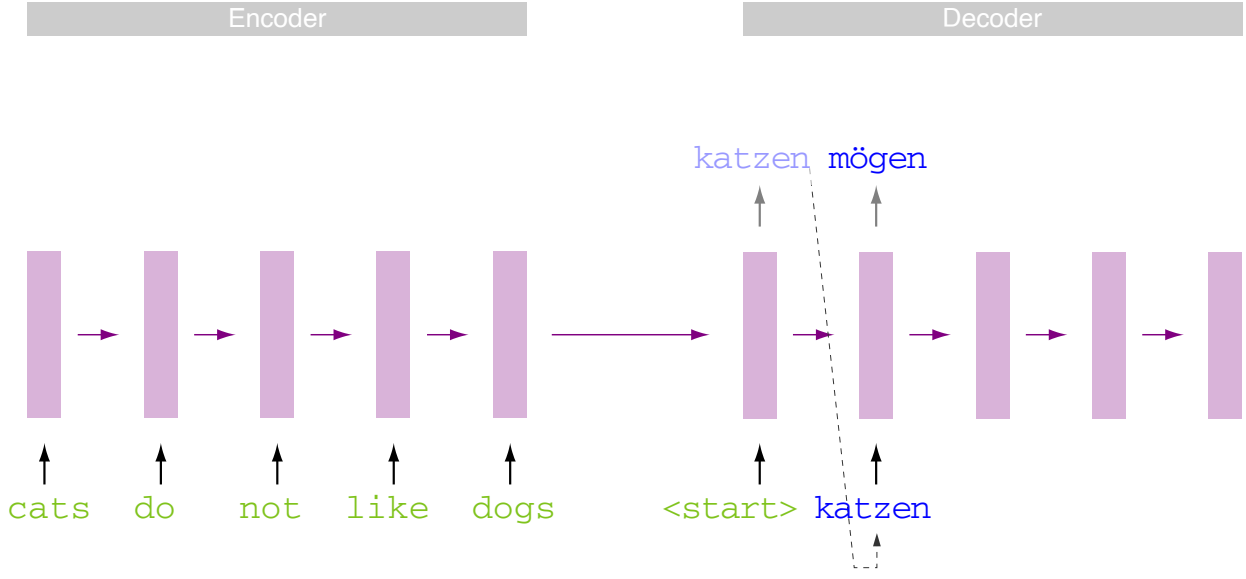


Neural language model (Bengio et al. 2000)

Recurrent neural language model with attention (Bahdanau et al. 2014)

1950 1960 1970 1980 1990 1995 2000 2002 2004 2006 2008 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026

A neural language model
tackles the probability maximization via loss minimization.

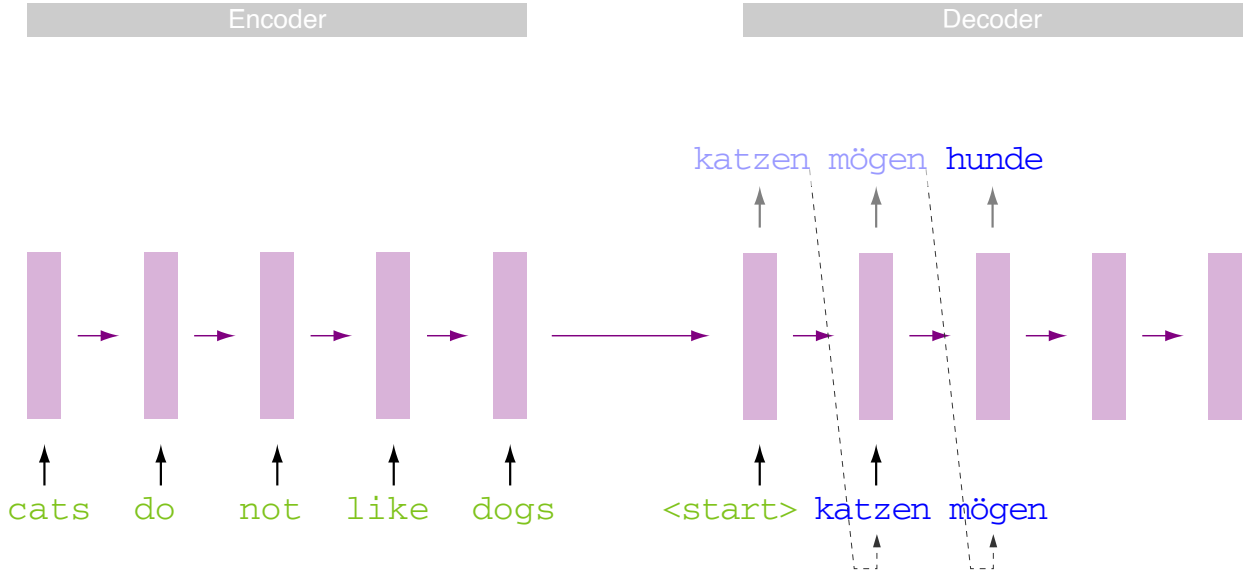


Neural language model (Bengio et al. 2000)

Recurrent neural language model with attention (Bahdanau et al. 2014)

1950 1960 1970 1980 1990 1995 2000 2002 2004 2006 2008 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026

A neural language model
tackles the probability maximization via loss minimization.

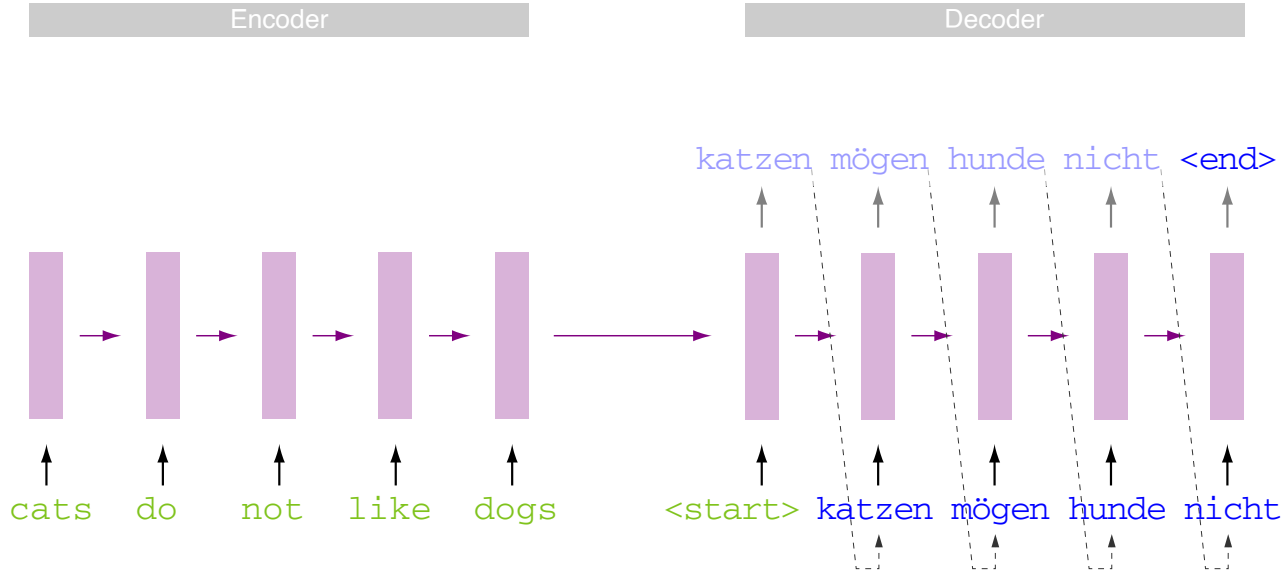


Neural language model (Bengio et al. 2000)

Recurrent neural language model with attention (Bahdanau et al. 2014)

1950 1960 1970 1980 1990 1995 2000 2002 2004 2006 2008 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026

A neural language model
tackles the probability maximization via loss minimization.

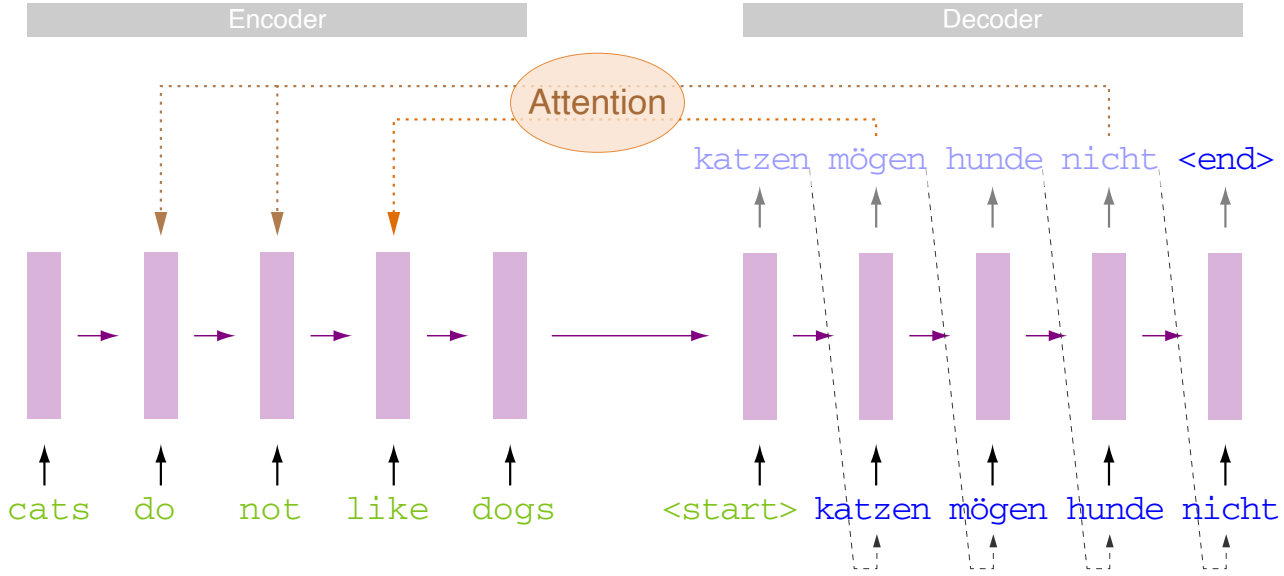


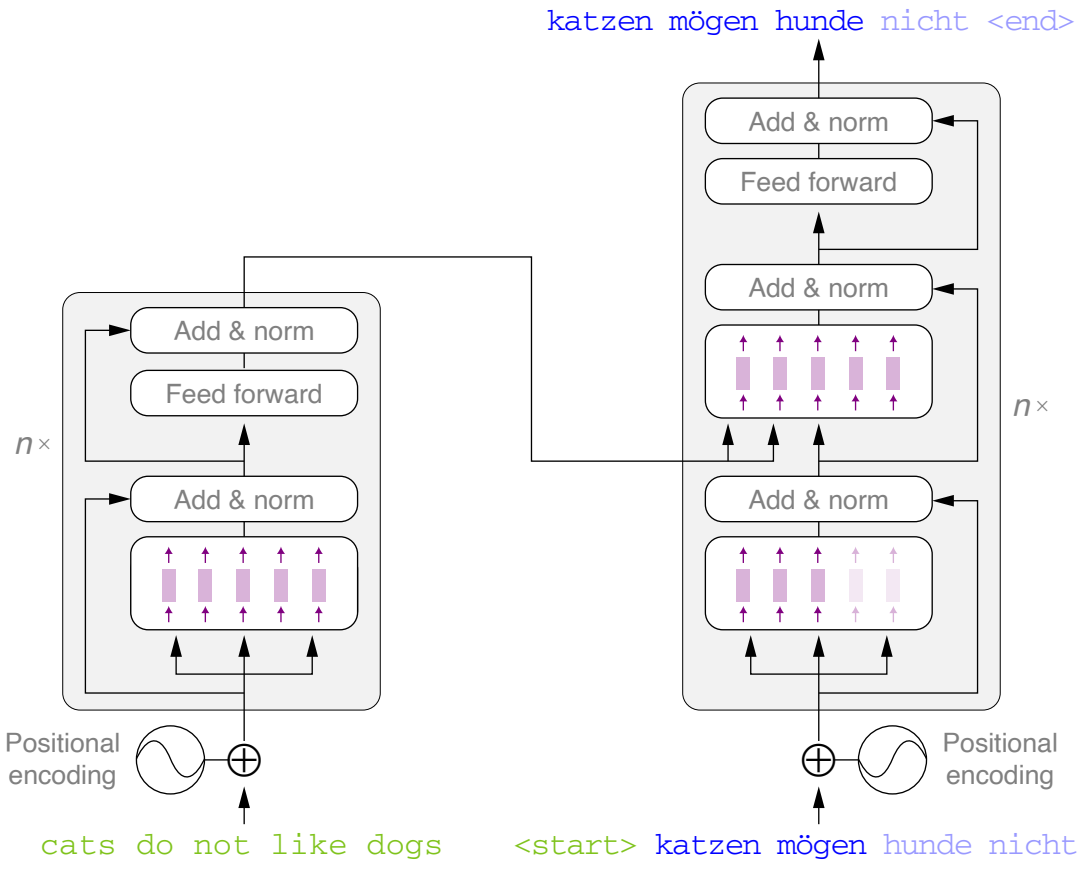
Neural language model (Bengio et al. 2000)

Recurrent neural language model with attention (Bahdanau et al. 2014)

1950 1960 1970 1980 1990 1995 2000 2002 2004 2006 2008 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026

A neural language model
tackles the probability maximization via loss minimization.

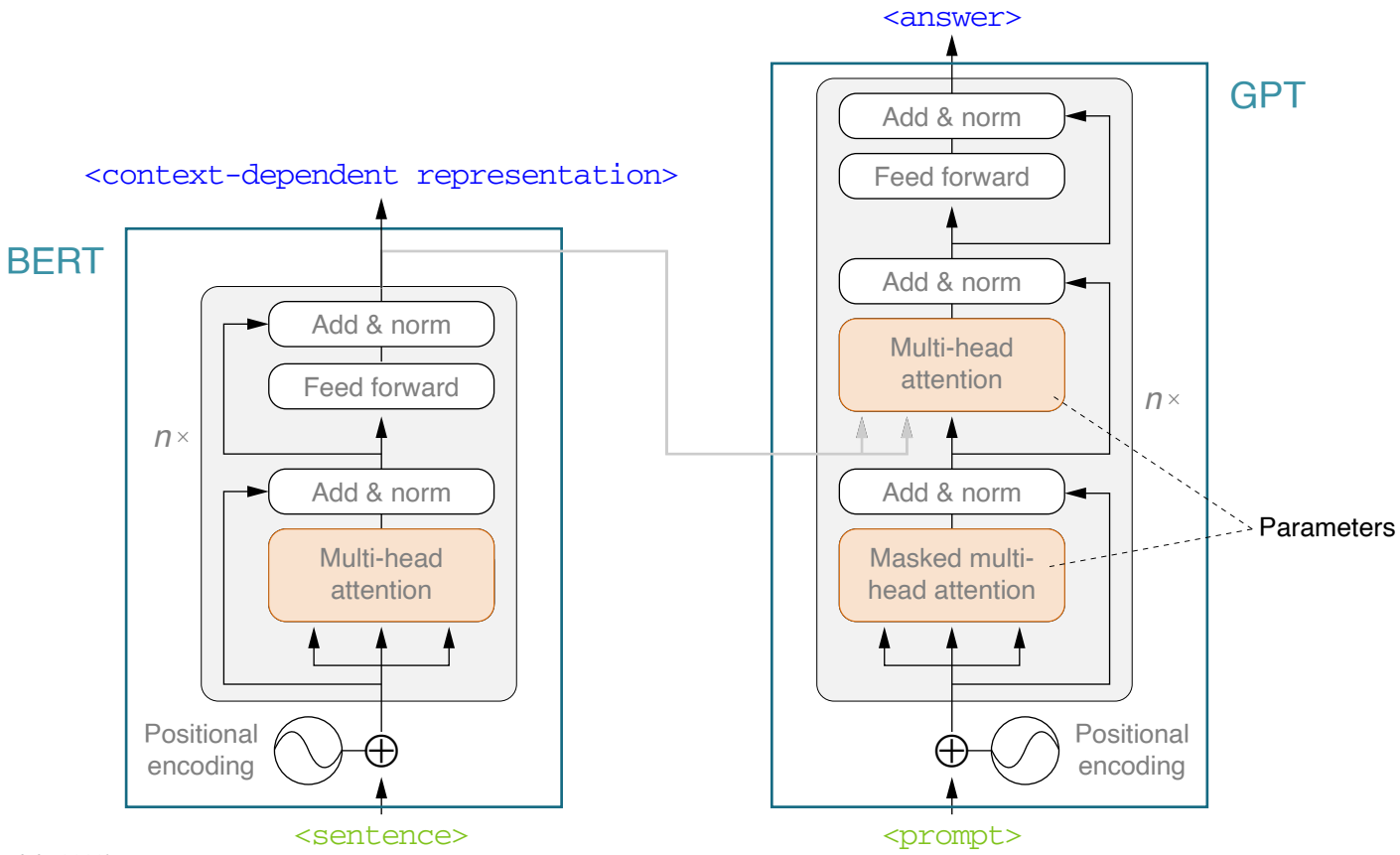



katzen mögen hunde nicht <end>
Add & norm
Feed forward
Add & norm
Add & norm
n ×
Add & norm
Feed forward
Add & norm
n ×
Positional encoding
Positional encoding
cats do not like dogs
<start> katzen mögen hunde nicht

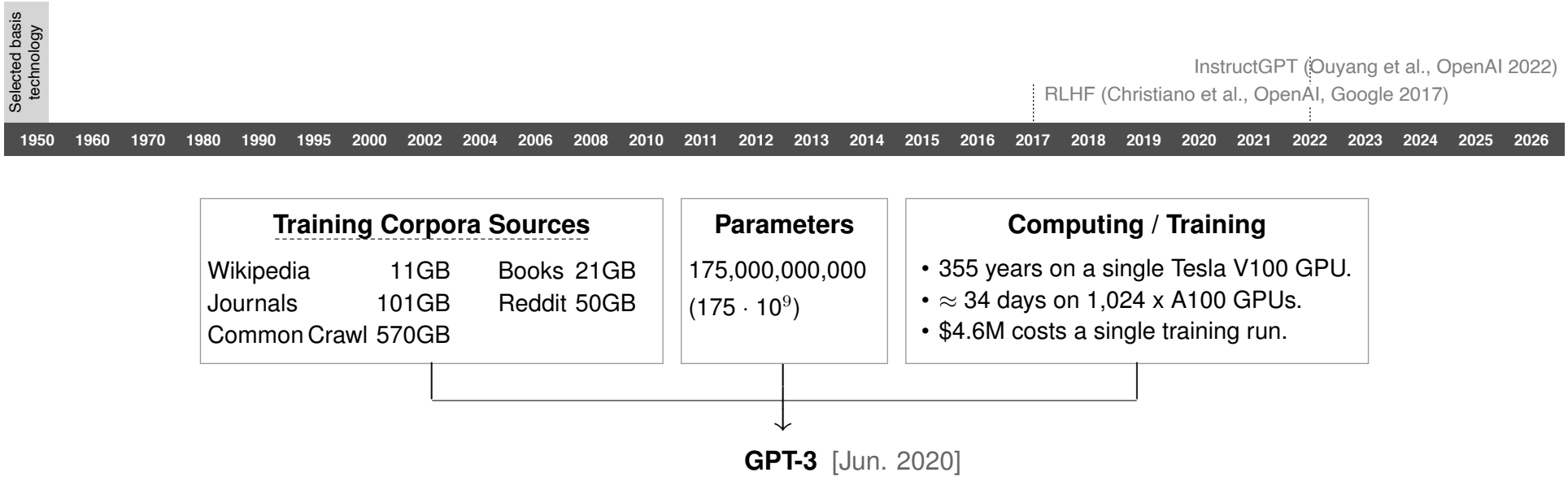
Selected basis technology

BERT (Devlin et al., Google 10/2018)
GPT (Radford et al., OpenAI 6/2018)

1950 1960 1970 1980 1990 1995 2000 2002 2004 2006 2008 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026



Transformer models catalog (Amatriain 2023)



Transformer models catalog (Amatriain 2023)

Training Corpora Sources

Wikipedia	11GB	Books	21GB
Journals	101GB	Reddit	50GB
Common Crawl	570GB		

Parameters

175,000,000,000
($175 \cdot 10^9$)

Computing / Training

- 355 years on a single Tesla V100 GPU.
- $\approx$ 34 days on 1,024 x A100 GPUs.
- $4.6M costs a single training run.

↓

GPT-3 [Jun. 2020]

\+ Learn to follow instructions and to comply with answer policies.

(1) Fine-tuning of GPT-3 to follow instructions: 13,000 popular prompts with hand-written answers.

(2) Training of a reward model: 33,000 prompts with 4-9 answers, ranked from best to worse.

(3) Training of the fine-tuned GPT-3 model from Step (1) to follow the reward policy.

↓

GPT-3.5 (InstructGPT) [Jan. 2022]

Training Corpora Sources

Wikipedia	11GB	Books	21GB
Journals	101GB	Reddit	50GB
Common Crawl	570GB		

Parameters

175,000,000,000
$(175 \cdot 10^9)$

Computing / Training

- 355 years on a single Tesla V100 GPU.
- $\approx$ 34 days on 1,024 x A100 GPUs.
- $4.6M costs a single training run.

↓

GPT-3 [Jun. 2020]

\+ Learn to follow instructions and to comply with answer policies.

(1) Fine-tuning of GPT-3 to follow instructions: 13,000 popular prompts with hand-written answers.

(2) Training of a reward model: 33,000 prompts with 4-9 answers, ranked from best to worse.

(3) Training of the fine-tuned GPT-3 model from Step (1) to follow the reward policy.

↓

GPT-3.5 (InstructGPT) [Jan. 2022]

\+ Fine-tuning of GPT-3.5 to comply with even stricter guardrails.

↓

ChatGPT [Nov. 2022]

Transformer models catalog (Amatriain 2023)

Training Corpora Sources

Wikipedia	11GB	Books	21GB
Journals	101GB	Reddit	50GB
Common Crawl	570GB		

Parameters

175,000,000,000
$(175 \cdot 10^9)$

Computing / Training

- 355 years on a single Tesla V100 GPU.
- $\approx$ 34 days on 1,024 x A100 GPUs.
- $4.6M costs a single training run.

↓

GPT-3 [Jun. 2020]

\+ Learn to follow instructions and to comply with answer policies.

(1) Fine-tuning of GPT-3 to follow instructions: 13,000 popular prompts with hand-written answers.
(2) Training of a reward model: 33,000 prompts with 4-9 answers, ranked from best to worse.
(3) Training of the fine-tuned GPT-3 model from Step (1) to follow the reward policy.

↓

GPT-3.5 (InstructGPT) [Jan. 2022]

\+ Fine-tuning of GPT-3.5 to comply with even stricter guardrails.

↓

ChatGPT [Nov. 2022]

Generative AI Authorship Verification

PAN 2024

Lab on Digital Text Forensics and Stylometry

pan.webis.de pan@webis.de

“Voight-Kampff” Generative AI Authorship Verification

Janek Bevendorff
Matti Wiegmann
Jussi Karlgren
Luise Dürlich
Evangelia Gogoulou
Aarne Talman
Efstathios Stamatatos
Martin Potthast
Benno Stein

Leipzig University
Bauhaus-Universität Weimar
Silo AI
RISE Research Institutes of Sweden
University of Helsinki
University of the Aegean
University of Kassel
hessian.AI
ScaDS.AI

Voight-Kampff* Generative AI Authorship Verification

Given two texts, one written by a human, the other by a large language model: decide which text was written by whom.

* From the 1982 science fiction film *Blade Runner*. The Voight-Kampff is a polygraph-like machine used by blade runners to determine whether an individual is a replicant. [Wikipedia]

Voight-Kampff* Generative AI Authorship Verification

Given two texts, one written by a human, the other by a large language model: decide which text was written by whom.

Task variants	Allowed assignment patterns
1. { ? , ? }	1. { A , 🤖 }
2. { ? , ? }	2. { A , 🤖 }, { A , A }
3. { ? , ? }	3. { A , 🤖 }, { 🤖 , 🤖 }
4. { ? , ? }	4. { A , 🤖 }, { A , A }, { 🤖 , 🤖 }
5. { ? , ? }	5. { A , 🤖 }, { A , A }, { A , B }
6. { ? , ? }	6. { A , 🤖 }, { A , A }, { A , B }, { 🤖 , 🤖 }
7. ?	7. A , 🤖

A, B, 🤖 represent texts from human authors A, B, and an LLM respectively. Increasing difficulty from 1 to 7.

* From the 1982 science fiction film *Blade Runner*. The Voight-Kampff is a polygraph-like machine used by blade runners to determine whether an individual is a replicant. [Wikipedia]

Voight-Kampff* Generative AI Authorship Verification

Given two texts, one written by a human, the other by a large language model: decide which text was written by whom.

Task variants	Allowed assignment patterns
1. { ? , ? }	1. { A , 🤖 }
2. { ? , ? }	2. { A , 🤖 }, { A , A }
3. { ? , ? }	3. { A , 🤖 }, { 🤖 , 🤖 }
4. { ? , ? }	4. { A , 🤖 }, { A , A }, { 🤖 , 🤖 }
5. { ? , ? }	5. { A , 🤖 }, { A , A }, { A , B }
6. { ? , ? }	6. { A , 🤖 }, { A , A }, { A , B }, { 🤖 , 🤖 }
7. ?	7. A , 🤖

A, B, 🤖 represent texts from human authors A, B, and an LLM respectively. Increasing difficulty from 1 to 7.

* From the 1982 science fiction film *Blade Runner*. The Voight-Kampff is a polygraph-like machine used by blade runners to determine whether an individual is a replicant. [Wikipedia]

Voight-Kampff* Generative AI Authorship Verification

Given two texts, one written by a human, the other by a large language model: decide which text was written by whom.

Task variants	Allowed assignment patterns
1. { ? , ? }	1. { A , 🤖 }
2. { ? , ? }	2. { A , 🤖 }, { A , A }
3. { ? , ? }	3. { A , 🤖 }, { 🤖 , 🤖 }
4. { ? , ? }	4. { A , 🤖 }, { A , A }, { 🤖 , 🤖 }
5. { ? , ? }	5. { A , 🤖 }, { A , A }, { A , B }
6. { ? , ? }	6. { A , 🤖 }, { A , A }, { A , B }, { 🤖 , 🤖 }
7. ?	7. A , 🤖

A, B, 🤖 represent texts from human authors A, B, and an LLM respectively. Increasing difficulty from 1 to 7.

* From the 1982 science fiction film *Blade Runner*. The Voight-Kampff is a polygraph-like machine used by blade runners to determine whether an individual is a replicant. [Wikipedia]

Voight-Kampff* Generative AI Authorship Verification

Given two texts, one written by a human, the other by a large language model: decide which text was written by whom.

Task variants	Allowed assignment patterns
1. { ?, ? }	1. { A, 🤖 }
2. { ?, ? }	2. { A, 🤖 }, { A, A }
3. { ?, ? }	3. { A, 🤖 }, { 🤖, 🤖 }
4. { ?, ? }	4. { A, 🤖 }, { A, A }, { 🤖, 🤖 }
5. { ?, ? }	5. { A, 🤖 }, { A, A }, { A, B }
6. { ?, ? }	6. { A, 🤖 }, { A, A }, { A, B }, { 🤖, 🤖 }
7. ?	7. A, 🤖

A, B, 🤖 represent texts from human authors A, B, and an LLM respectively. Increasing difficulty from 1 to 7.

* From the 1982 science fiction film *Blade Runner*. The Voight-Kampff is a polygraph-like machine used by blade runners to determine whether an individual is a replicant. [Wikipedia]

Generative AI Authorship Verification (dataset creation)

- Human text: 1,359 US news articles from 2021, crawled from Google News.
- Article decomposition with ChatGPT 4. Prompt excerpt:
  - `"Summarize the key points in 10 bullet points."`
  - `"Classify the article type ('breaking news', 'government agency statement', ..."`
  - `"Determine the article's target audience ('general public', 'children', ..."`
  - `"Classify whether the article's stance is 'left-leaning', ..."`

Generative AI Authorship Verification (dataset creation)

- Human text: 1,359 US news articles from 2021, crawled from Google News.
- Article decomposition with ChatGPT 4. Prompt excerpt:
  - `"Summarize the key points in 10 bullet points."`
  - `"Classify the article type ('breaking news', 'government agency statement', ..."`
  - `"Determine the article's target audience ('general public', 'children', ..."`
  - `"Classify whether the article's stance is 'left-leaning', ..."`
- Machine text: reconstruction of articles by 13 LLMs. Prompt excerpt:

  `"You are a journalist writing {{ article_type }}.  In your article, cover the following ..."`

Generative AI Authorship Verification (dataset creation)

- Human text: 1,359 US news articles from 2021, crawled from Google News.
- Article decomposition with ChatGPT 4. Prompt excerpt:
  - `"Summarize the key points in 10 bullet points."`
  - `"Classify the article type ('breaking news', 'government agency statement', ..."`
  - `"Determine the article's target audience ('general public', 'children', ..."`
  - `"Classify whether the article's stance is 'left-leaning', ..."`
- Machine text: reconstruction of articles by 13 LLMs. Prompt excerpt:

  `"You are a journalist writing {{ article_type }}. In your article, cover the following ..."`
- The generated texts are cleaned manually of artifacts.
- Test data: 3,411 pairs of human and machine text.
- Test data variants to analyze selected robustness aspects:

  unicode obfuscation, cropped text (35 words), cross-topic pairs, cross-language pairs

Generative AI Authorship Verification (baselines and submissions)

- 13 Baseline systems, among others:
  - DetectGPT [Mitchell et al., 2023]
  - Fast-DetectGPT [Bao et al., 2023]
  - DetectLLM LRR and NPR [Su et al., 2023]
  - Binoculars [Hans et al., 2024]
- Evaluation measures:

  ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1, Mean of all
- 30 Submissions
- Winning system:

  ensemble of Binoculars and a fine-tuned Mistral + Llama
- Popular approaches:

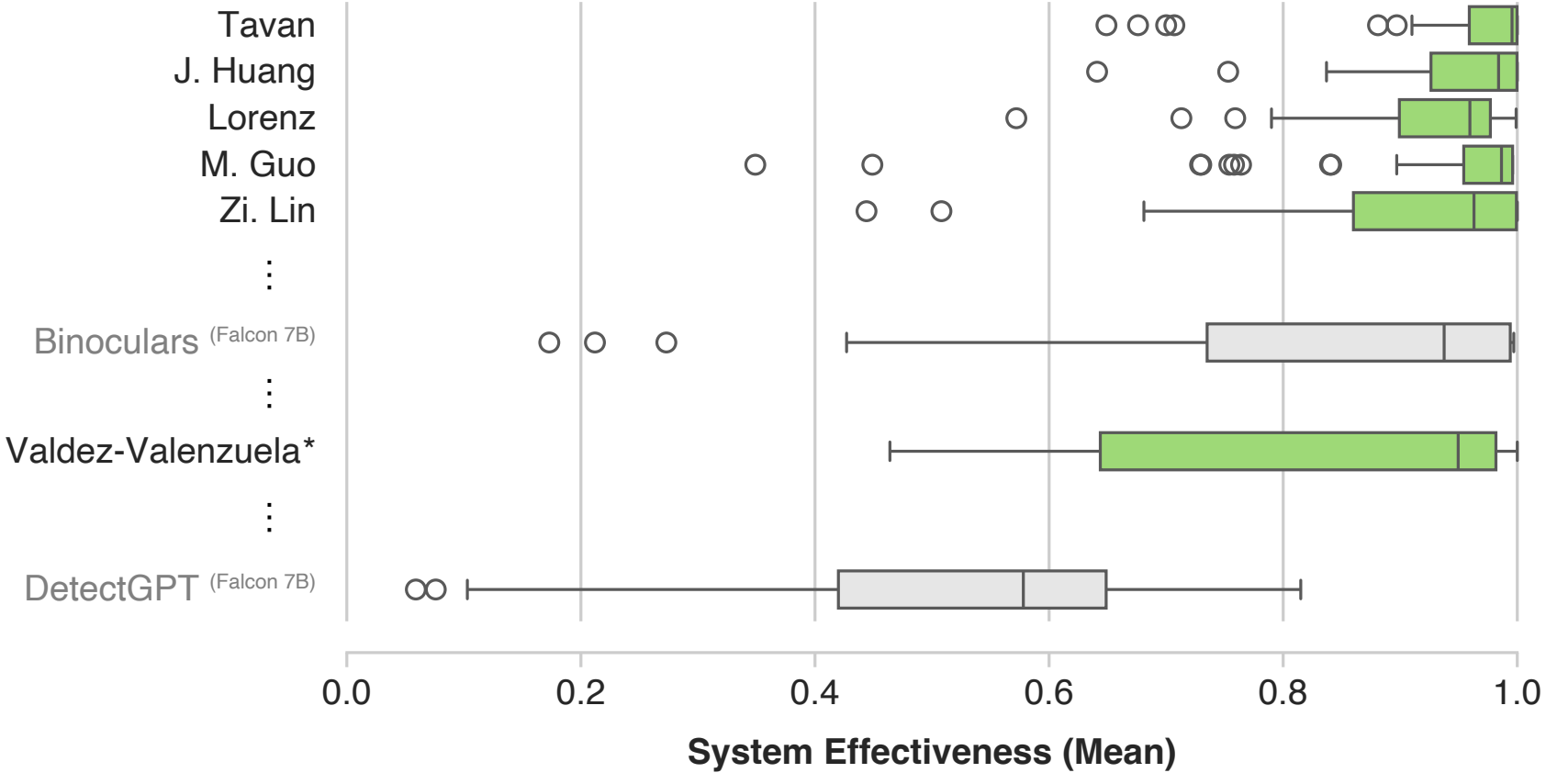
  fine-tuned BERT (20), perplexity (11), stylometry (5), ensembles (5), augmented data (6)

Generative AI Authorship Verification (systems ranking)

	Team	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
1	Tavan	**0.961**	**0.928**	0.912	0.884	**0.932**	**0.924**
2	J. Huang	0.931	0.926	**0.928**	**0.905**	0.913	0.921
3	Lorenz	0.925	0.869	0.882	0.875	0.869	0.886
4	M. Guo	0.889	0.875	0.887	0.884	0.884	0.884
5	Zi. Lin	0.851	0.850	0.850	0.852	0.849	0.851
			⋮				
Baseline Binoculars (Falcon 7B)		0.751	0.780	0.734	0.720	0.720	0.741
			⋮				
14	Valdez-Valenzuela	0.741*	0.760*	0.718*	0.711*	0.695*	0.727*
			⋮				
Baseline DetectGPT (Falcon 7B)		0.409	0.526	0.425	0.413	0.412	0.439

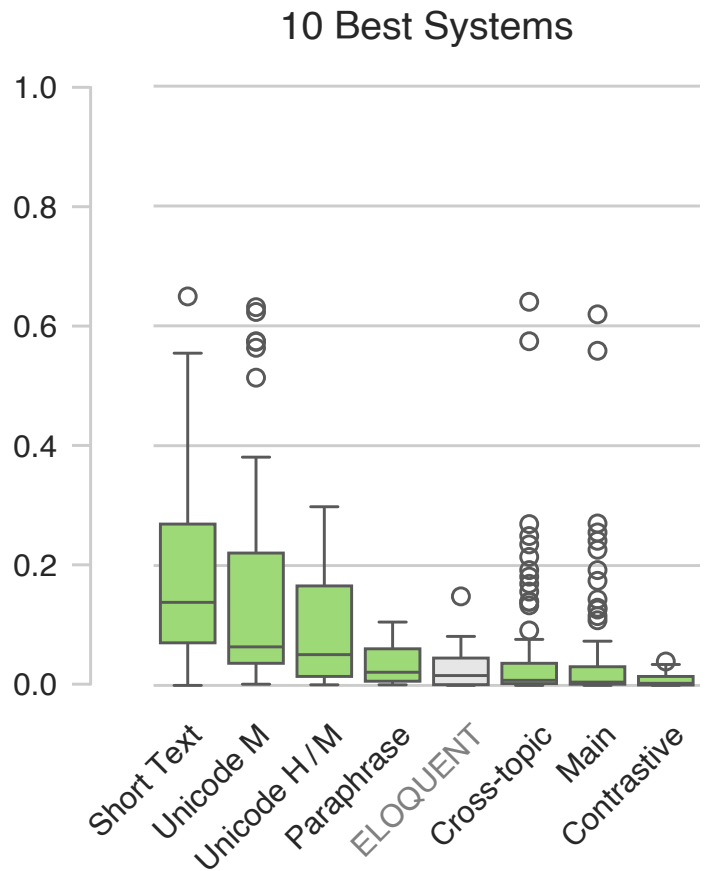
* Scores estimated due to run failures on short texts.

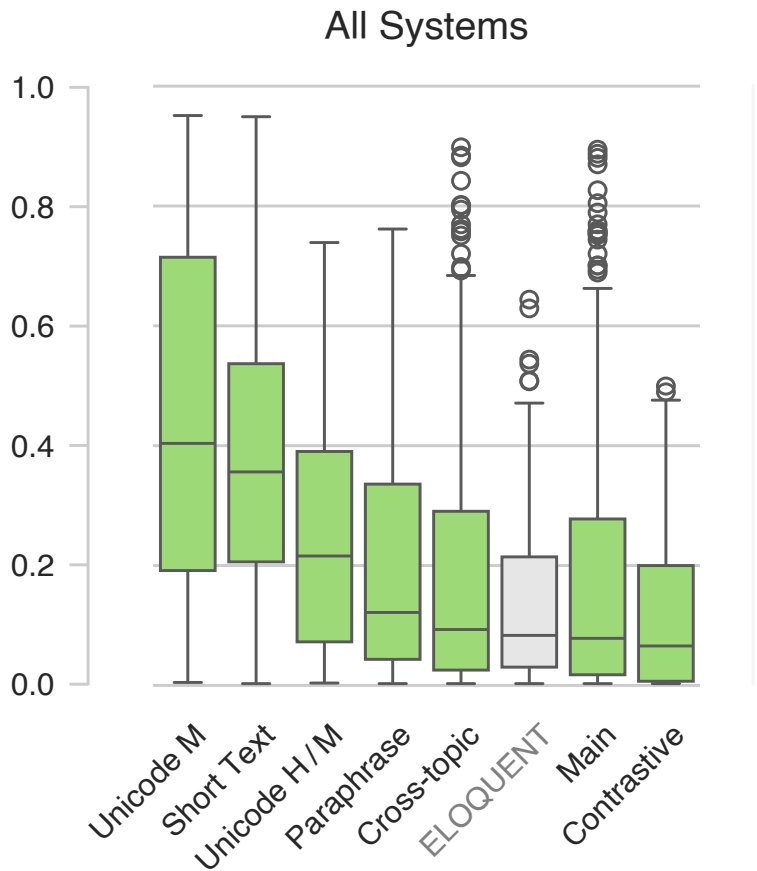
Generative AI Authorship Verification (score distribution)



* Scores estimated due to run failures on short texts.

Generative AI Authorship Verification (dataset difficulty as $1-$effectiveness)





AI-related Research at Webis

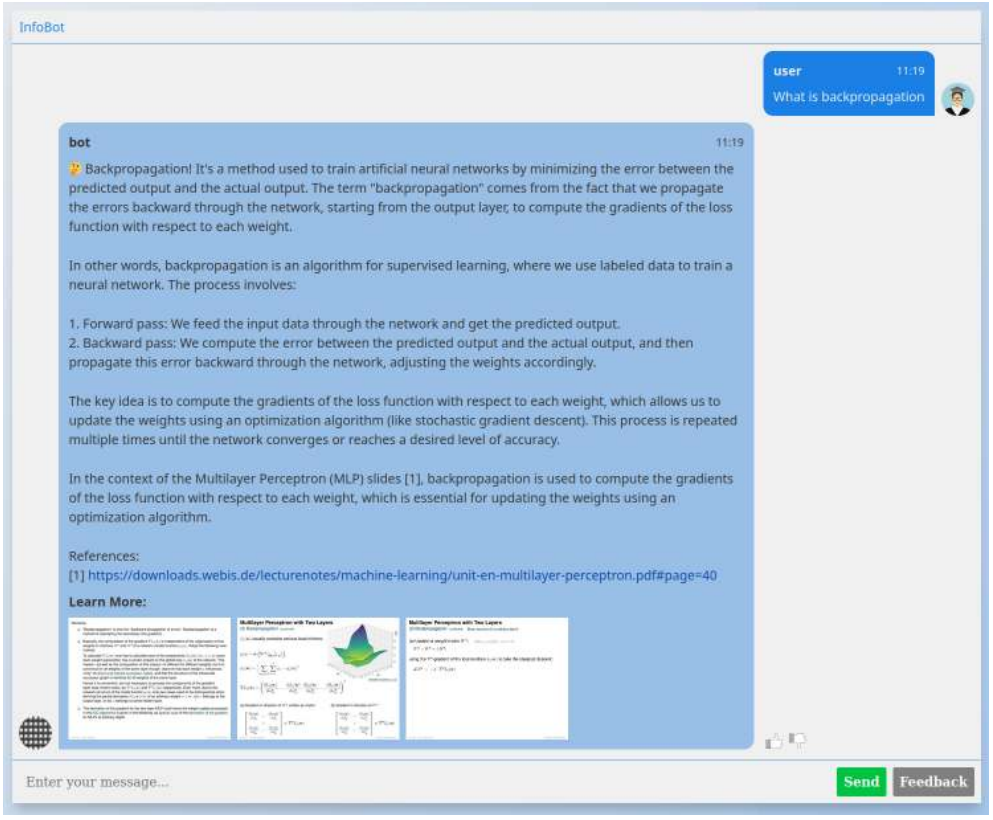
AI in Authorship Analytics.	LLM detection
AI in ML Education.	The InfoBot project
AI in the Humanties.	Automatic discourse generation
AI in Media Design.	Futuring Machines
AI in NLP research.	Identifying the values behind arguments
AI in Political Sciences teaching.	The SKILL project
AI in Social Sciences research.	Curating social media feeds
AI in Web Search.	Retrieval augmented generation


Multilayer Perceptron Basics
(1) Overcoming the Linear Separability Restriction
A minimum multilayer perceptron

Benno 1.0

Benno 2.0

The InfoBot Project

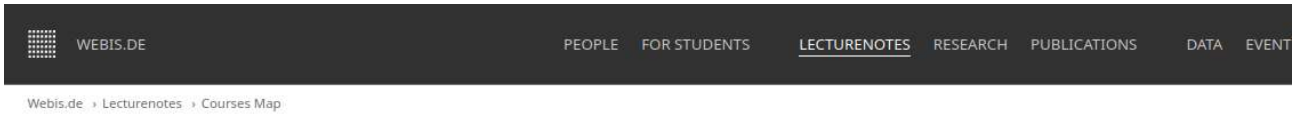


<InfoBotURL@webis.de>

- exploit own teaching resources
  → recognize formalization dialectics
- consider all Webis courses
  → show impact on related fields
- combine slides with explanations
  → show additional connections
  → provide the best entry points
- consider dialog context
  → allow for followup question
- learning theory perspective
  - encourage to draw conclusions
  - consider individual prior knowledge
  - construct individual mental model

The InfoBot Project (resources)

lecturenotes.webis.de



Webis.de › Lecturenotes › Courses Map

Courses Map

The table below organizes the Webis courses (see copyright), which cover relevant contents from our research areas. Clicking a table cell will bring you to the respective course slides. criterion to see at which universities or at what level our courses are taught.

Click here to filter...

Algorithmen und Datenstrukturen	Einführung	Algorithm Engineering	Sortieren	Datenstrukturen	Suchen	Graphalgorithmen			
Datenbanken	Einführung	Konzeptueller Datenbank-entwurf	Logischer Datenbank-entwurf	Grundlagen relationaler Anfragesprachen	SQL	Entwurfstheorie relationaler Datenbanken	Physischer Datenbank-entwurf		
Web-Technologie	Einführung	Kommunikation und Protokolle für Web-Systeme	Dokument-sprachen	Server-Technologien	Client-Technologien	Architekturen und Middleware	Semantic Web		
Information Retrieval	Introduction	Indexing	Retrieval Models	User Interface	Evaluation	IR Applications			
Natural Language Processing	Introduction	Corpus Linguistics	Text Models	Language Models	Words	Syntax	Semantics	Discourse	Bias and Fairness
	NLP Applications								
Wahrscheinlich-keitstheorie und Statistik	Einführung	Wahrscheinlich-keitsbegriff	Kombinatorik	Bedingte Wahrschein-lichkeit	Zufallsgrößen und Maßzahlen	Die Binomial-verteilung	Das Gesetz der großen Zahlen	Die Normalverteilung	Hypothesentests
Machine		Machine Learning			Support Vector				

The InfoBot Project (resources)

lecturenotes.webis.de

WEBIS.DE PEOPLE FOR STUDENTS LECTURENOTES RESEARCH PUBLICATIONS DATA EVENT

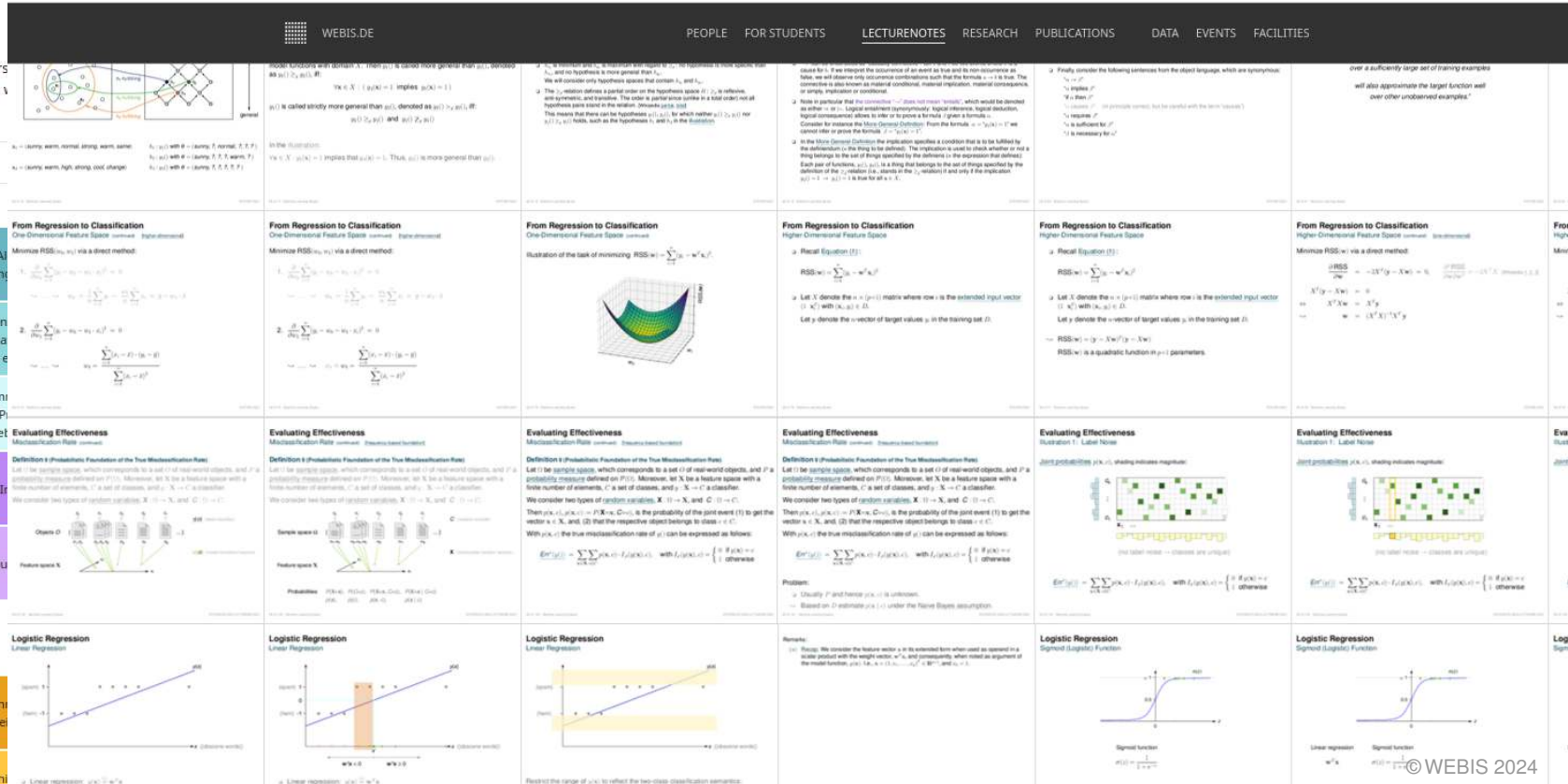
Webis.de › Lecturenotes › Courses Map

Courses Map

The table below organizes the Webis cours
criterion to see at which universities or at

Click here to filter...

Algorithmen und Datenstrukturen	Einführung	Al Eng
Datenbanken	Einführung	Kon Da e
Web-Technologie	Einführung	Komm und Pl Wet
Information Retrieval	Introduction	Ir
Natural Language Processing	Introduction	Corpu
	NLP Applications	
Wahrscheinlich-keitstheorie und Statistik	Einführung	Wahr ke
Machine		Machi



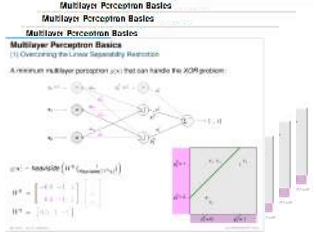
The InfoBot Project (search engine index)

```
\begin{bsslide}
\small
Remarks (prior probability model)
\hypertarget{prior-probability-mo
\begin{itemize}
\setlength{\itemsep}{1ex}
\item
In the example it is presumed th
\begin{itemize}
\setlength{\itemsep}{0.5ex}
\item
the set of diagnoses is complete
\item
that $A_1$ and $A_2$ are mutuall
```

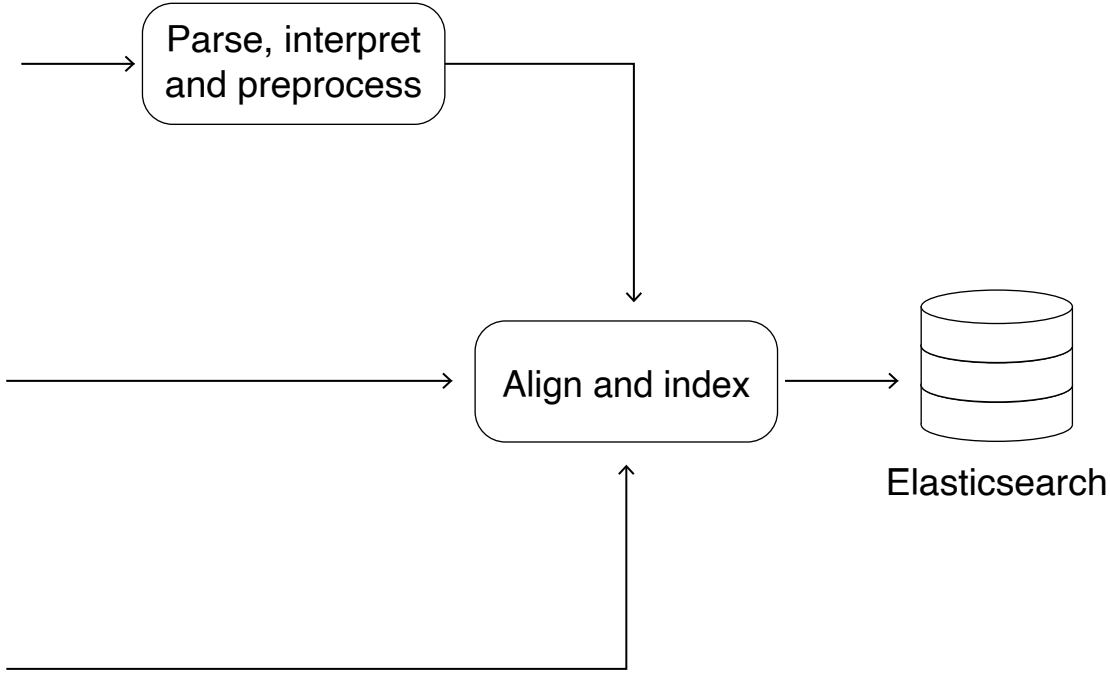
Latex sources

```
%%% NOTES.
%%%
%%% The prior probabilities ...
%%% (1) it accounts for...
%%% (2) it assigns prob...
%%% (3) these probabilities...
%%% ...
```

Inline comments



Lecture slides



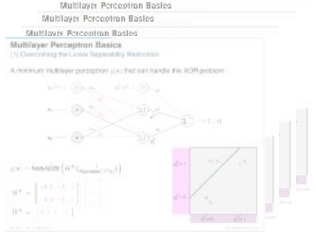
The InfoBot Project (search engine index)

```
\begin{bsslide}
\small
Remarks (prior probability model)
\hypertarget{prior-probability-mo
\begin{itemize}
\setlength{\itemsep}{1ex}
\item
In the example it is presumed th
\begin{itemize}
\setlength{\itemsep}{0.5ex}
\item
the set of diagnoses is complete
\item
that $A_1$ and $A_2$ are mutuall
```

Latex sources

```
%%% NOTES.
%%%
%%% The prior probabilities ...
%%% (1) it accounts for...
%%% (2) it assigns prob...
%%% (3) these probabilities...
%%% ...
```

Inline comments



Lecture slides

Parse, interpret and preprocess

Align and index

Elasticsearch

Query

"Ten blue links"

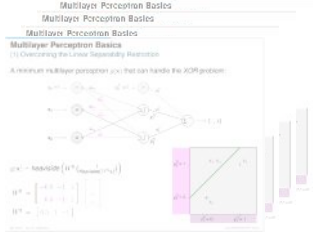
The InfoBot Project (search engine index)

```
\begin{bsslide}
\small
Remarks (prior probability model)
\hypertarget{prior-probability-mo
\begin{itemize}
\setlength{\itemsep}{1ex}
\item
In the example it is presumed th
\begin{itemize}
\setlength{\itemsep}{0.5ex}
\item
the set of diagnoses is complete
\item
that $A_1$ and $A_2$ are mutuall
```

Latex sources

```
%%% NOTES.
%%%
%%% The prior probabilities ...
%%% (1) it accounts for...
%%% (2) it assigns prob...
%%% (3) these probabilities...
%%% ...
```

Inline comments



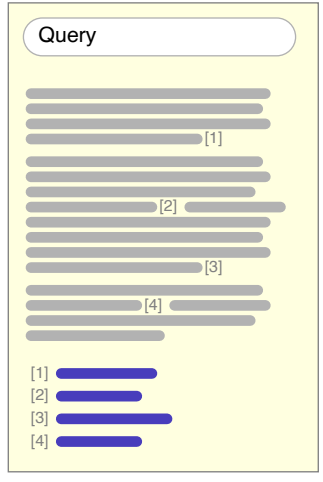
Lecture slides

Parse, interpret and preprocess

Align and index

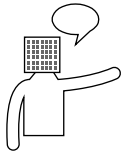
R etrieval
A ugemented
G eneration

Elasticsearch



"Ten blue links"

The InfoBot Project (dialog processing)

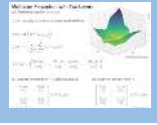


user 13:49

What is backprop-agation?

bot 13:50

Backpropagation is a method used in training artificial neural networks to calculate the gradients of ...



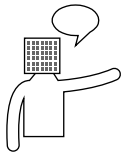
user 13:54

Is backpropogation gradient descend?

bot 13:55

Backpropagation and stochastic gradient descent (SGD) are two related but distinct concepts in machine learning...

The InfoBot Project (dialog processing)



user 13:49

What is backprop-agation?

bot 13:50

Backpropagation is a method used in training artificial neural networks to calculate the gradients of ...

user 13:54

Is backpropogation gradient descend?

bot 13:55

Backpropagation and stochastic gradient descent (SGD) are two related but distinct concepts in machine learning...

Web-interface

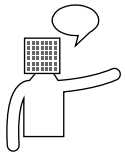
Infobot Server

Keyword Extractor

Elastic-search

LLM

The InfoBot Project (dialog processing)



user 13:49

What is backpropagation?

bot 13:50

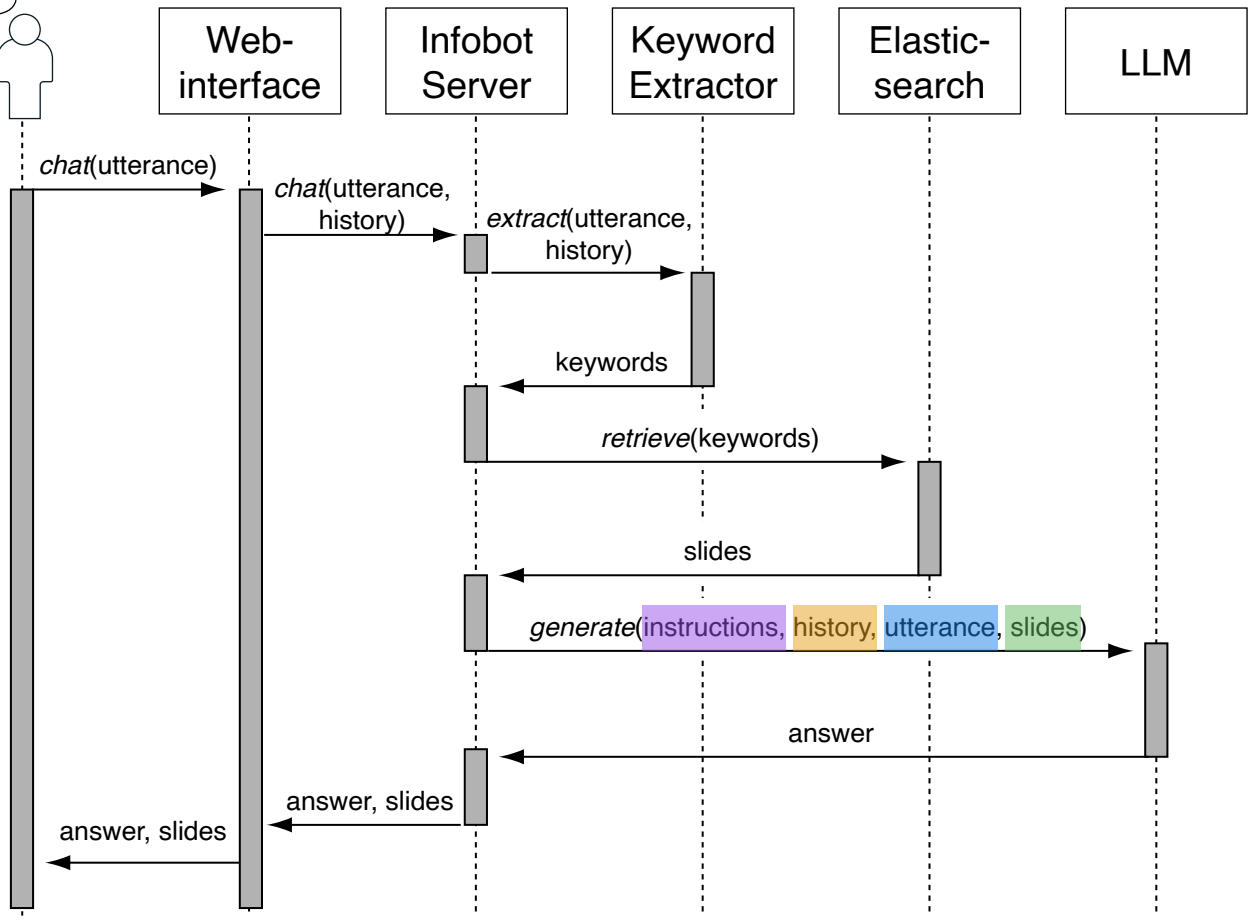
Backpropagation is a method used in training artificial neural networks to calculate the gradients of ...

user 13:54

Is backpropogation gradient descend?

bot 13:55

Backpropagation and stochastic gradient descent (SGD) are two related but distinct concepts in machine learning...



The InfoBot Project (instructions in the system prompt)

1. Behavioural instructions

   `"You are a friendly teaching assistant called 'Infobot' ..."`

2. Course information and URLs

   `"These are the courses taught by the Webis group ..."`

3. Citation instructions

   `"You should provide references to relevant slides when you are ..."`

4. Meta instructions

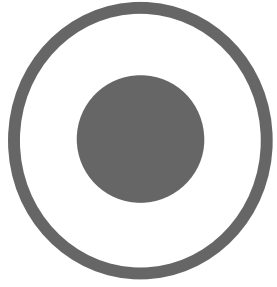
   `"Keep the answers short (maximum of two to three sentences) ..."`

5. Instructions for the retrieved slides (top three)

   `"Use the following information to construct your answer ..."`

The InfoBot Project (other specs of the RAG pipeline)

- Keyword extraction
  - KeyBERT (`all-mpnet-base-v2`)
  - word n-grams up to 5-grams
- Retrieval model
  - BM15 against slide title, subtitle and content
  - Reranking: weight BM15 result by keyword likelihood from KeyBERT
- Large language model
  - Meta Llama 3
  - 8 billion paramaters
  - 6-bit quantization



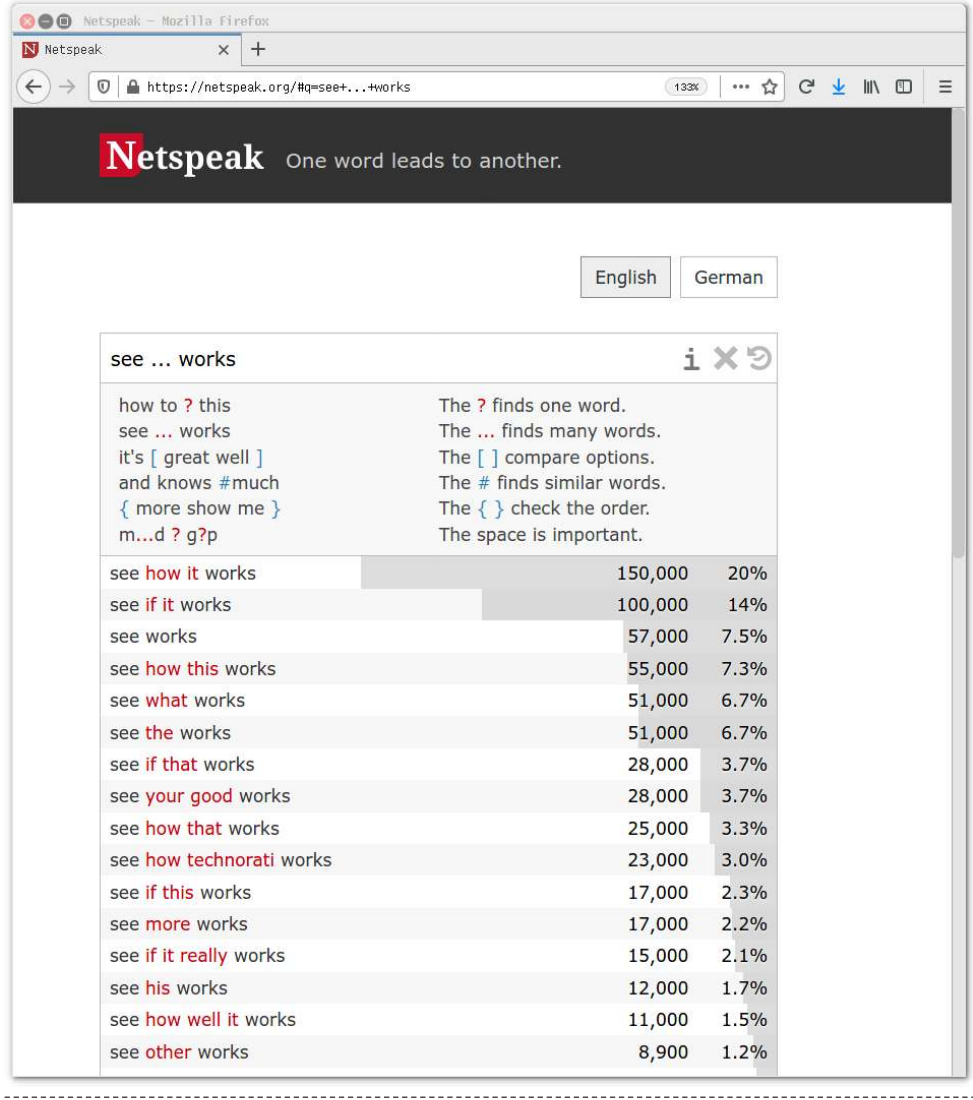
Wrapup

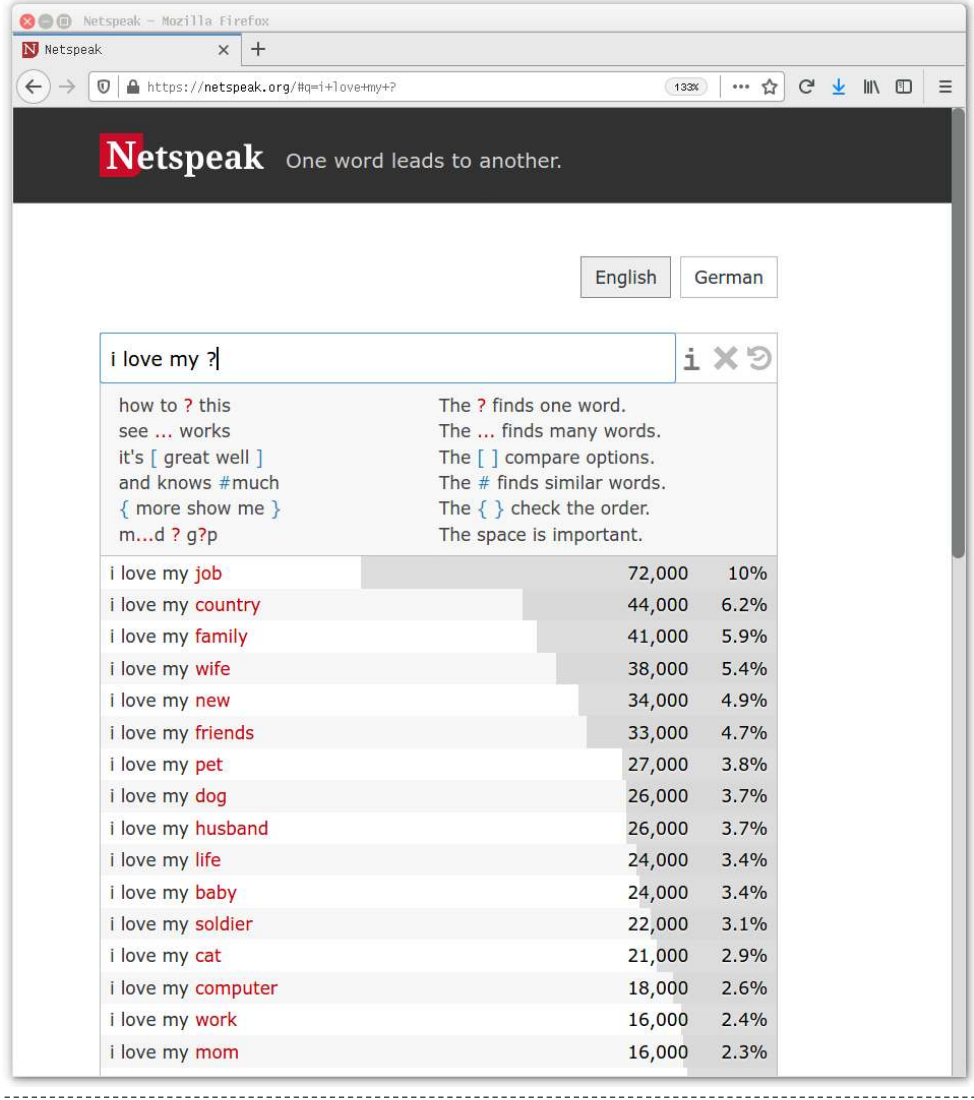
① About the Turing Test

② Some Background on Large Language Models

③ Generative AI Authorship Verification

④ The Infobot Project


Netspeak – Mozilla Firefox
Netspeak
https://netspeak.org/#q=see+...+works
133%
Netspeak One word leads to another.
English
German
see ... works
how to ? this
see ... works
it's [great well]
and knows #much
{ more show me }
m...d ? g?p
The ? finds one word.
The ... finds many words.
The [ ] compare options.
The # finds similar words.
The { } check the order.
The space is important.
see how it works 150,000 20%
see if it works 100,000 14%
see works 57,000 7.5%
see how this works 55,000 7.3%
see what works 51,000 6.7%
see the works 51,000 6.7%
see if that works 28,000 3.7%
see your good works 28,000 3.7%
see how that works 25,000 3.3%
see how technorati works 23,000 3.0%
see if this works 17,000 2.3%
see more works 17,000 2.2%
see if it really works 15,000 2.1%
see his works 12,000 1.7%
see how well it works 11,000 1.5%
see other works 8,900 1.2%


Netspeak – Mozilla Firefox
Netspeak
https://netspeak.org/#q=i+love+my+?
133%
Netspeak One word leads to another.
English
German
i love my ?
how to ? this
see ... works
it's [great well]
and knows #much
{ more show me }
m...d ? g?p
The ? finds one word.
The ... finds many words.
The [ ] compare options.
The # finds similar words.
The { } check the order.
The space is important.
i love my job 72,000 10%
i love my country 44,000 6.2%
i love my family 41,000 5.9%
i love my wife 38,000 5.4%
i love my new 34,000 4.9%
i love my friends 33,000 4.7%
i love my pet 27,000 3.8%
i love my dog 26,000 3.7%
i love my husband 26,000 3.7%
i love my life 24,000 3.4%
i love my baby 24,000 3.4%
i love my soldier 22,000 3.1%
i love my cat 21,000 2.9%
i love my computer 18,000 2.6%
i love my work 16,000 2.4%
i love my mom 16,000 2.3%