Large Language Models at Work

Retrieval, Authenticity, and Computational Social Science

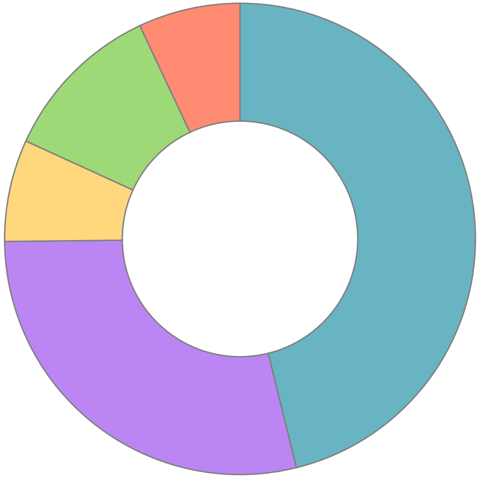
Benno Stein

Bauhaus-Universität
Weimar

Research Group Intelligent Information Systems [webis.de]

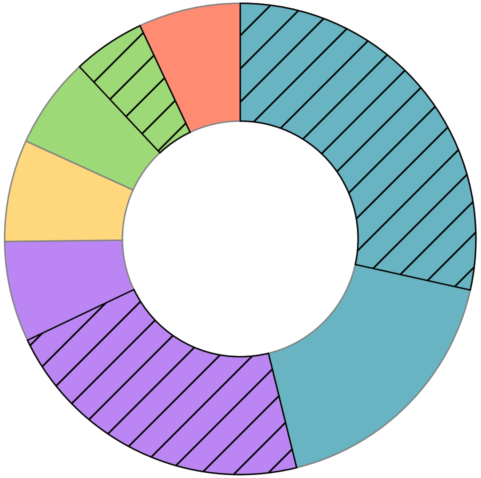
Webis Research – Technologies

Webis Research – Technologies



- 🟦 Information Retrieval
  - Ranking Paradigms
  - Evaluation and Benchmarking
- 🟪 Natural Language Processing
  - Algorithms
  - Corpus Curation
- 🟨 Data Mining and Machine Learning
  - Algorithms
  - Big Data Processing
- 🟩 Research Competitions
  - PAN Series
  - Touché Series
- 🟥 Platforms and Software
  - Automated Experiment Configuration and Execution

Webis Research – Technologies



▨ Use of LLM technology

▬ Information Retrieval
- Ranking Paradigms
- Evaluation and Benchmarking

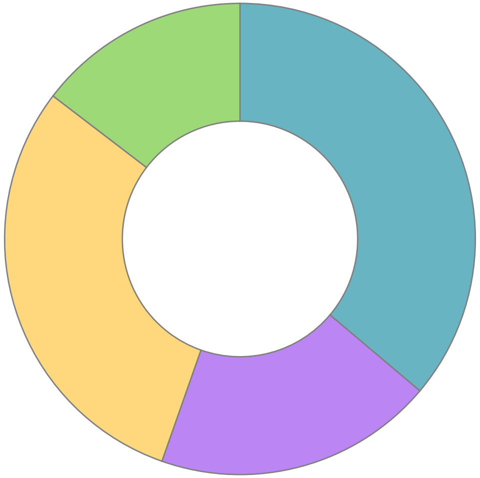
▬ Natural Language Processing
- Algorithms
- Corpus Curation

▬ Data Mining and Machine Learning
- Algorithms
- Big Data Processing

▬ Research Competitions
- PAN Series
- Touché Series

▬ Platforms and Software
- Automated Experiment Configuration and Execution

Webis Research – Applications



🟦 Web Search

- Search Engines: Chatnoir, Netspeak, PicaPica
- Conversational Search, RAG, Retrieval Axioms

🟪 Authorship Analytics and Provenance

- Author Identification and Obfuscation
- Text Watermarking

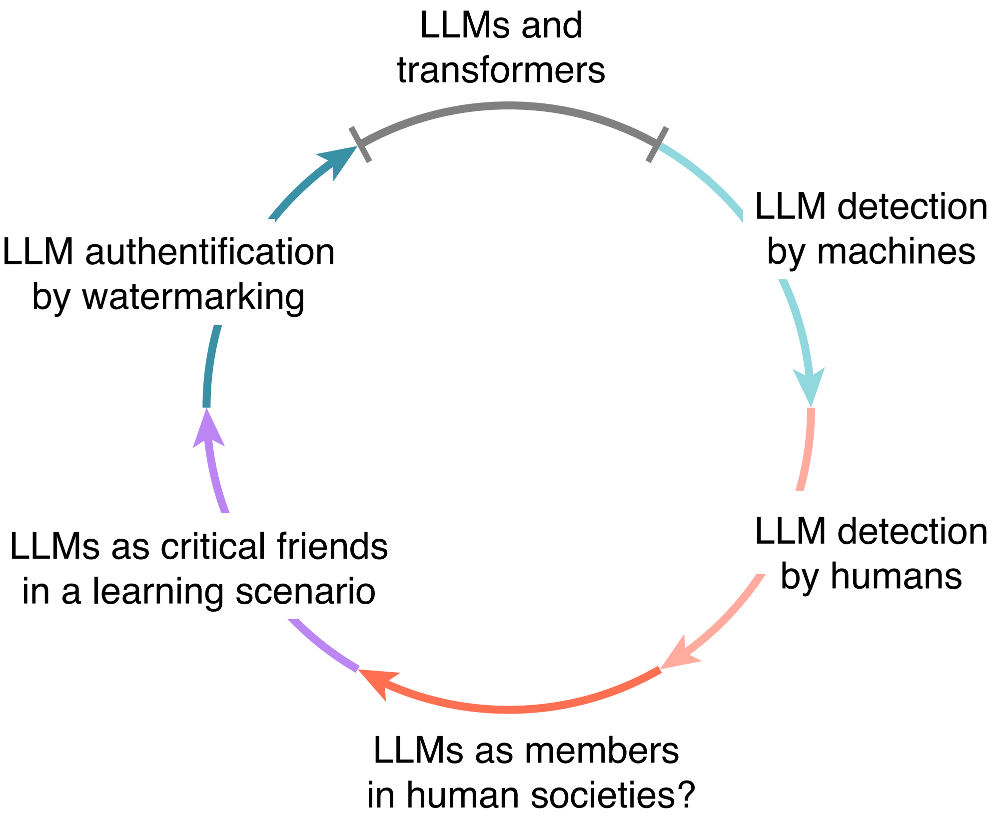
🟨 Computational Argumentation

- Argument Search: Args.me
- (multimodal, political) Argument Analytics

🟩 Social Media Analytics

- Information Nutrition Label
- Human Value Detection
- Trigger Warnings, Feed Analytics

Agenda for this Lecture



LLMs and transformers

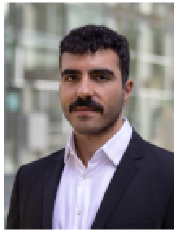
LLM detection by machines

LLM detection by humans

LLMs as members in human societies?

LLMs as critical friends in a learning scenario

LLM authentification by watermarking


Pierre Achkar
Leipzig


Christopher Akiki
Leipzig


Janek Bevendorff
Weimar


Niklas Deckers
Kassel

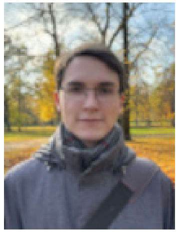

Theresa Elstner
Kassel


Maik Fröbe
Jena


Lukas Gienapp
Kassel


Marcel Gohsen
Weimar


Tim Gollub
Weimar


Tim Hagen
Kassel


Sebastian Heineking
Leipzig


Maximilian Heinrich
Weimar


Midhun Kanadan
Weimar


Jan Heinrich Merker
Jena

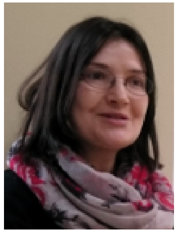

Nailia Mirzakhmedova
Weimar


Simon Ruth
Kassel


Ferdinand Schlatt
Jena

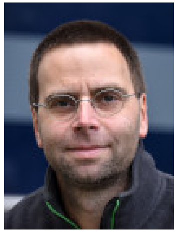

Michael Völske
Weimar


Matti Wiegmann
Weimar

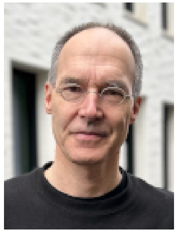

Magdalena Wolska
Weimar


Ines Zelch
Jena


Johannes Kiesel
GESIS


Matthias Hagen
Jena


Martin Potthast
Leipzig


Benno Stein
Weimar

Agenda

① Background on Large Language Models and Transformers

② Who is the Author? Generative LLM Authorship Verification

③ Turing X (interactive)

④ The Infobot Project – An LLM-based Teaching Prototype for Lectures

⑤ Watermarking Large Language Models

> *“You shall know a word by the company it keeps.”*
>
> [John Rupert Firth, 1957]

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

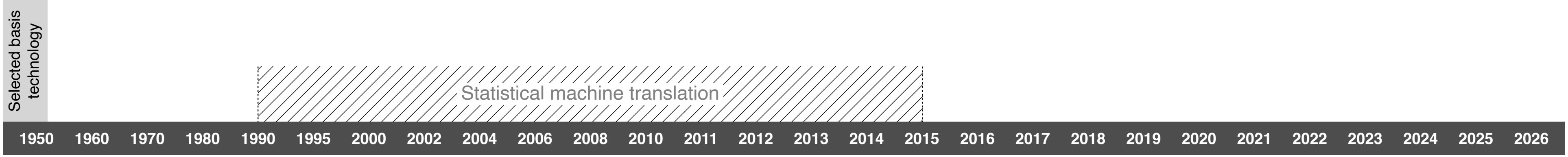
Example:

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

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

---

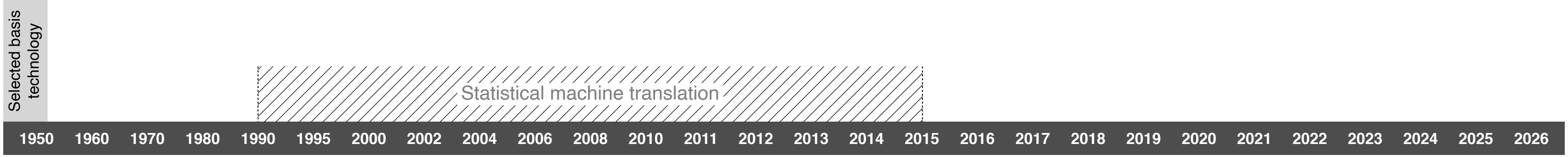
* Keyword: “Distributional Semantics” – Key players: J. R. Firth, Zellig S. Harris, in the 1950s



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

(1) `i love my ?` N N

(2) `see ... works.` N N



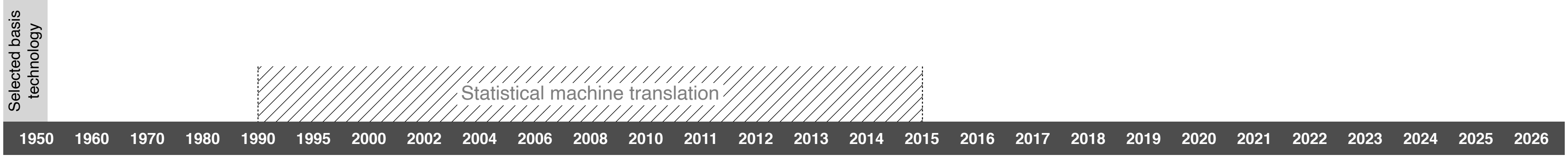
> A statistical language model
> is a probability distribution over all possible texts.

(1) `i love my ?` N N

(2) `see ... works.` N N

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.`

Sentence translation means *probability maximization* :

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

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

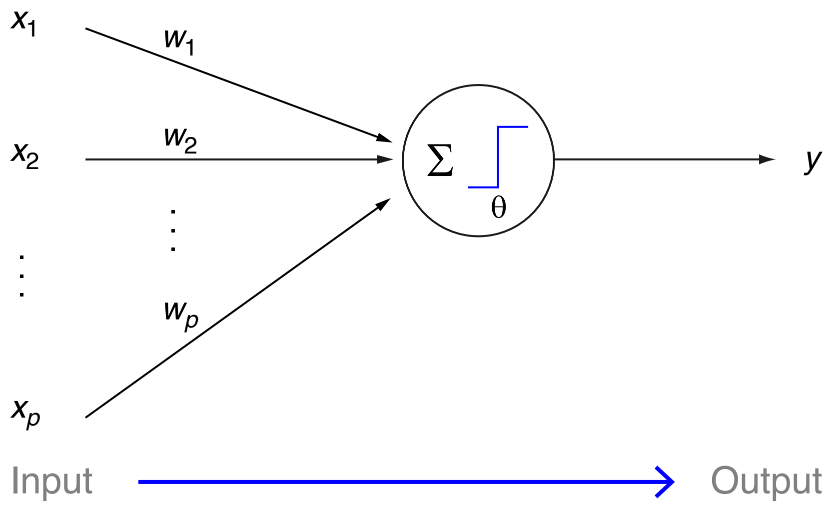
$$p(\texttt{ich habe keine katze} \mid \texttt{i love my cat})$$

Selected basis technology

Feedforward Neural Network (implementation of single perceptron, Rosenblatt 1958)

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.



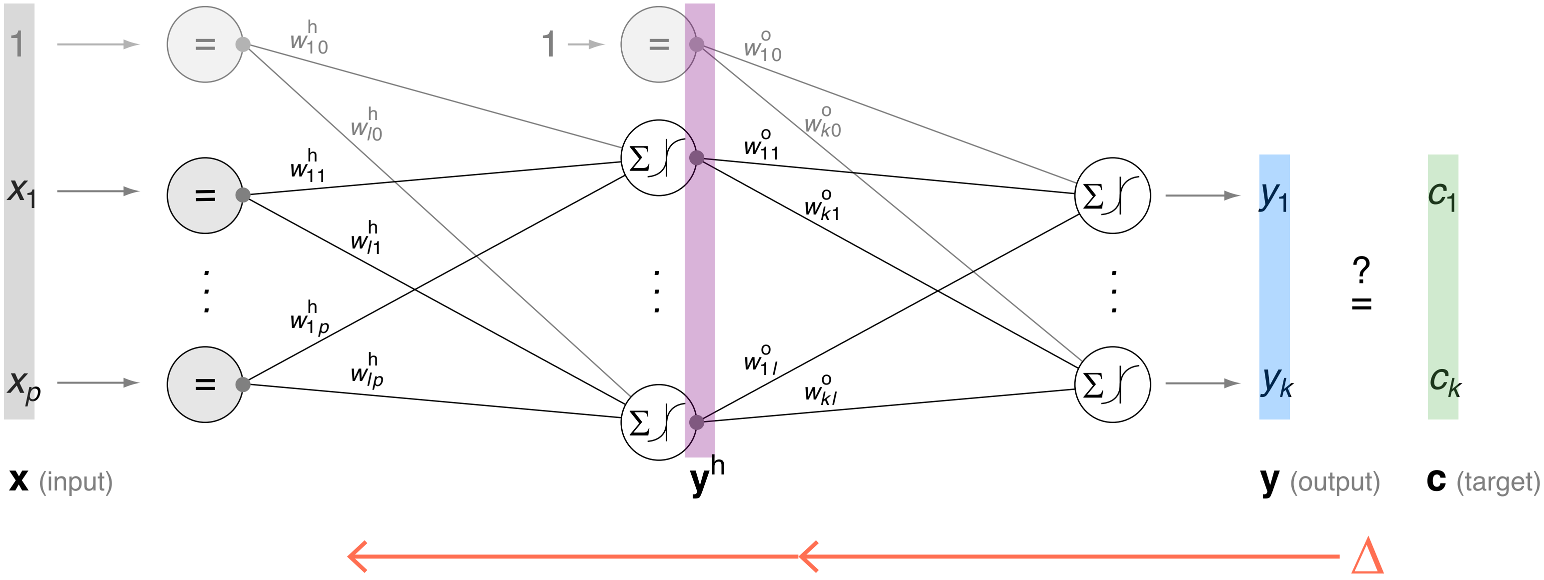
Selected basis technology

Multilayer Perceptron with backpropagation (Werbos 1982, Rumelhart 1982)

Backpropagation with automatic differentiation (Linnainmaa 1970)

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.

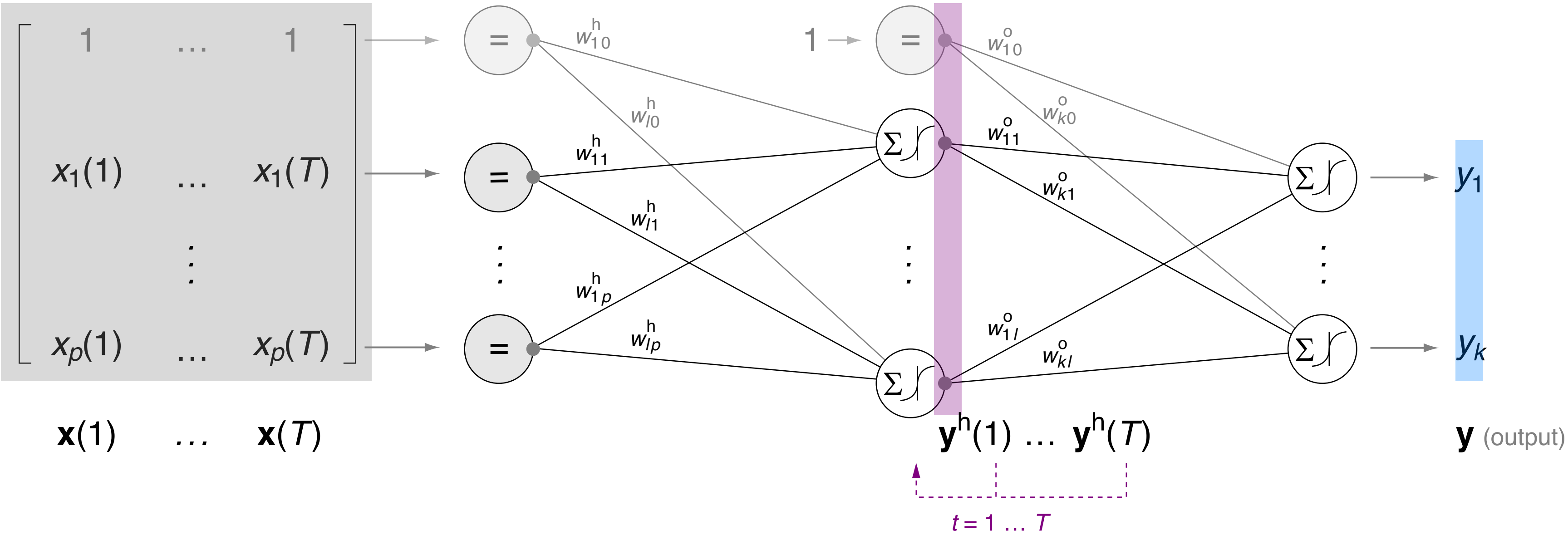


Selected basis technology

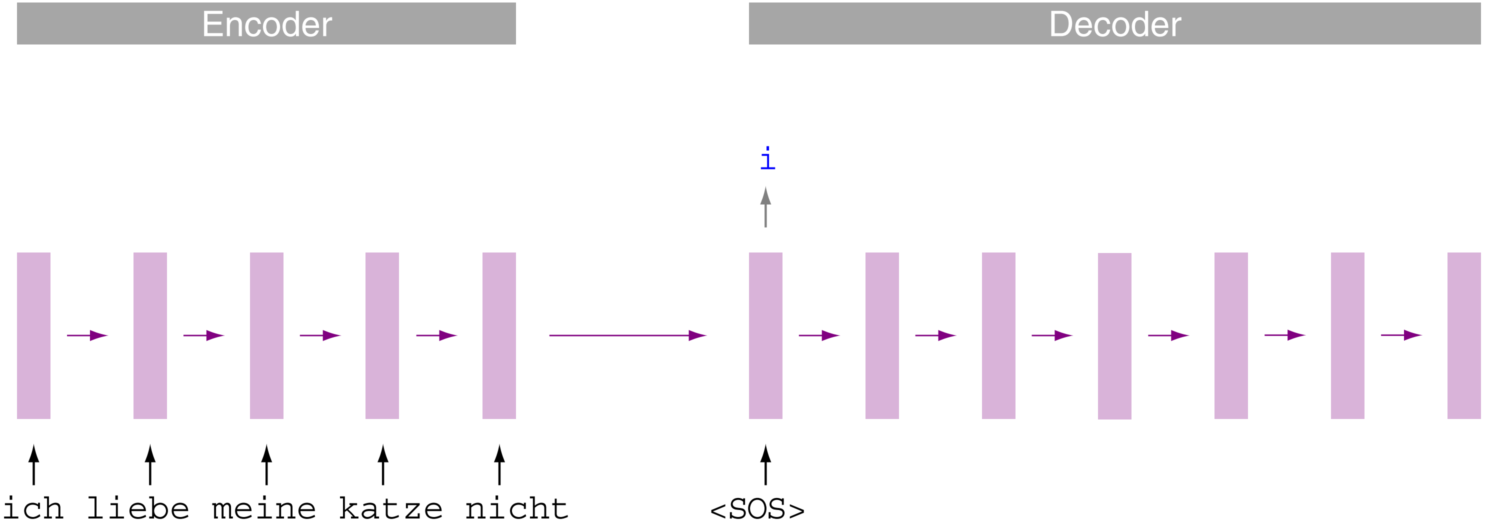
Recurrent Neural Network (Hopfield 1982)

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.



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



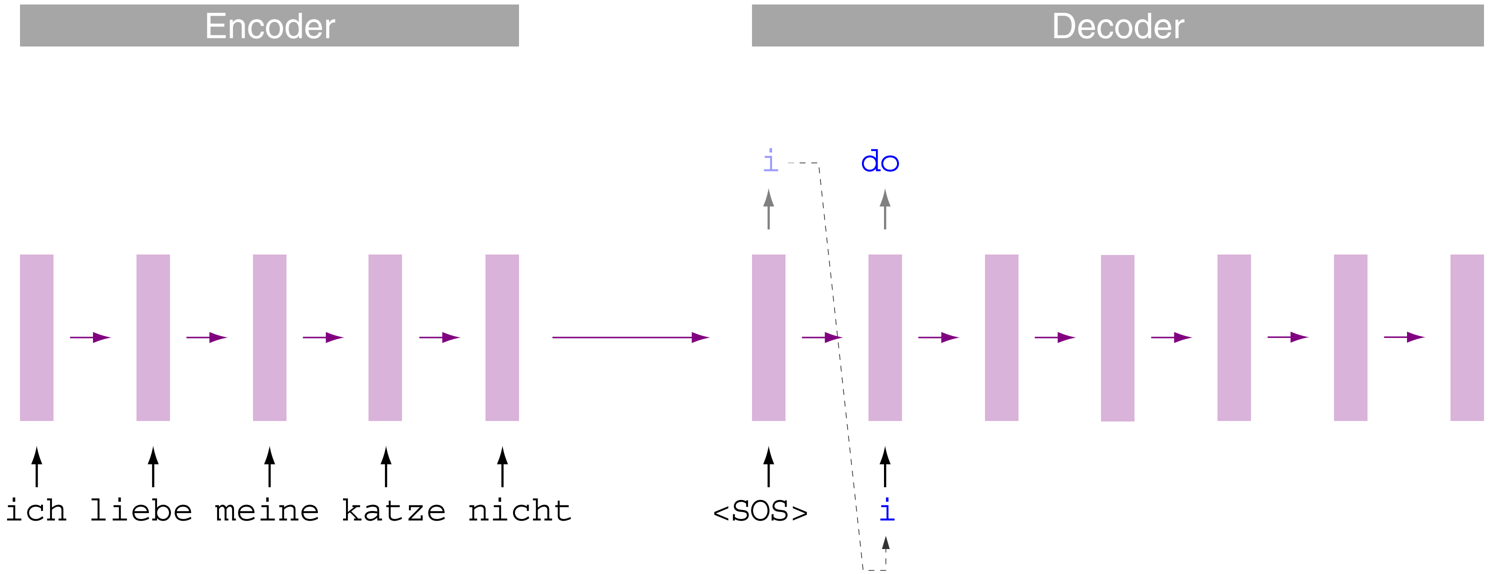
Selected basis technology

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.



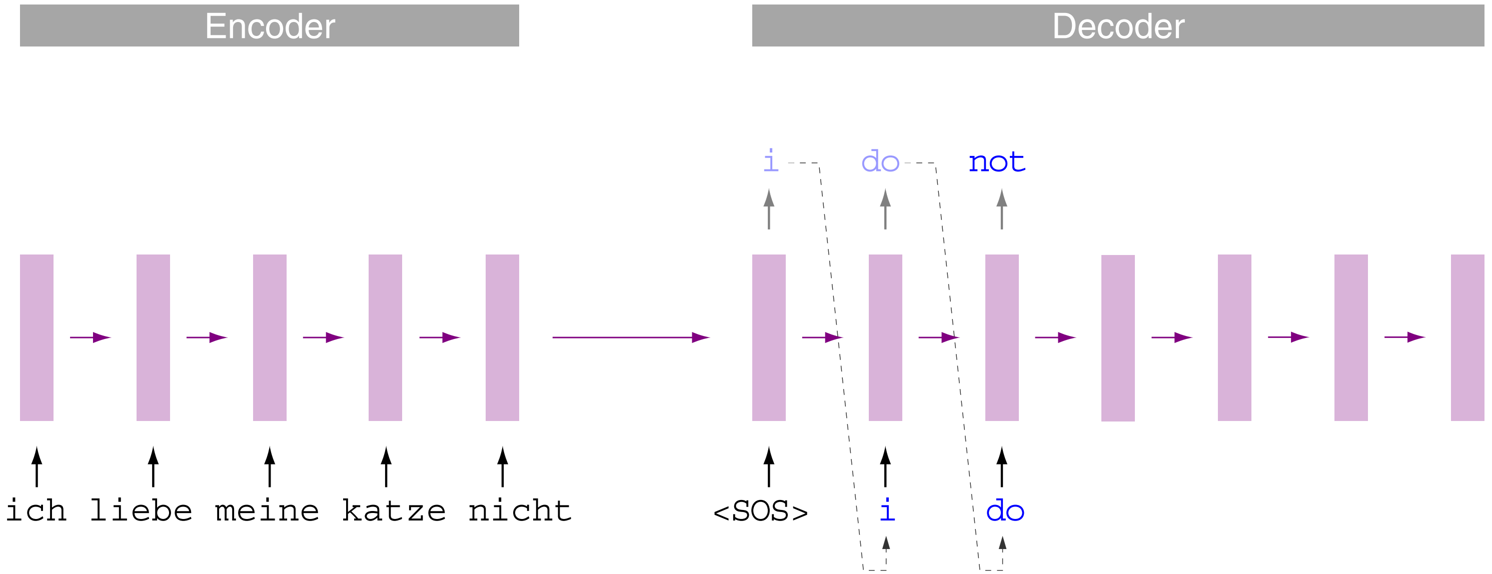
Selected basis technology

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.



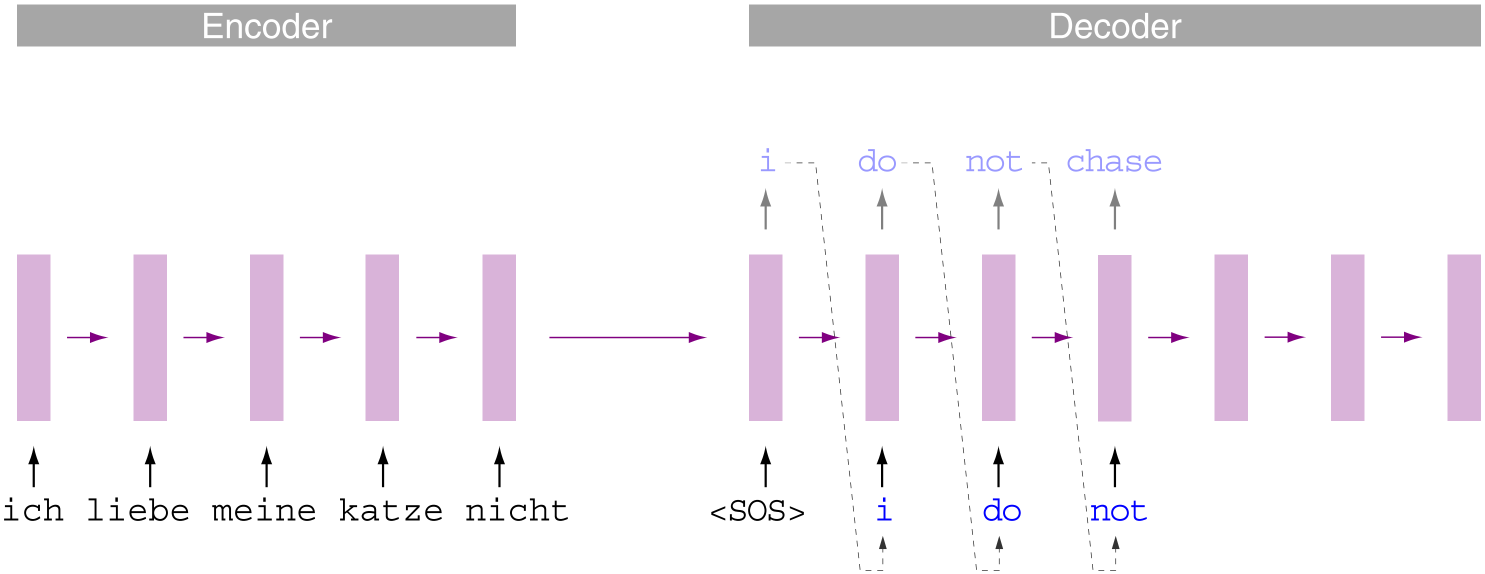
Selected basis technology

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.



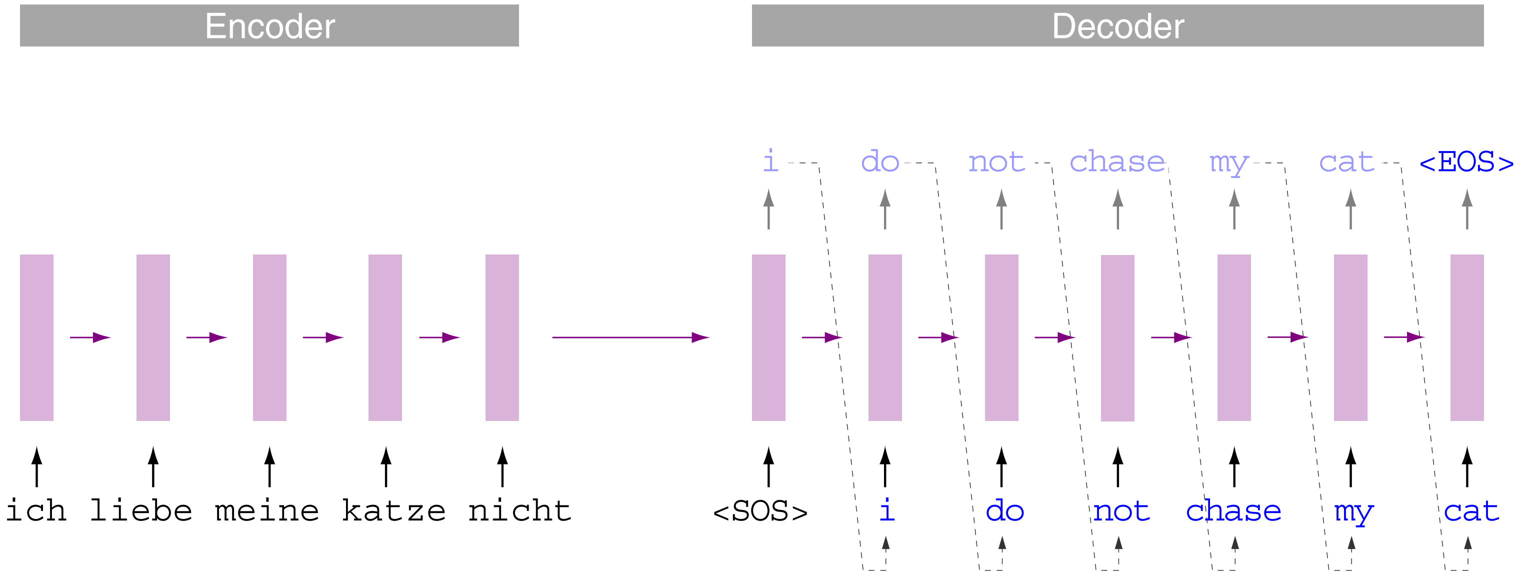
Selected basis technology

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.



Encoder

Decoder

i do not chase my cat <EOS>

ich liebe meine katze nicht

<SOS> i do not chase my cat

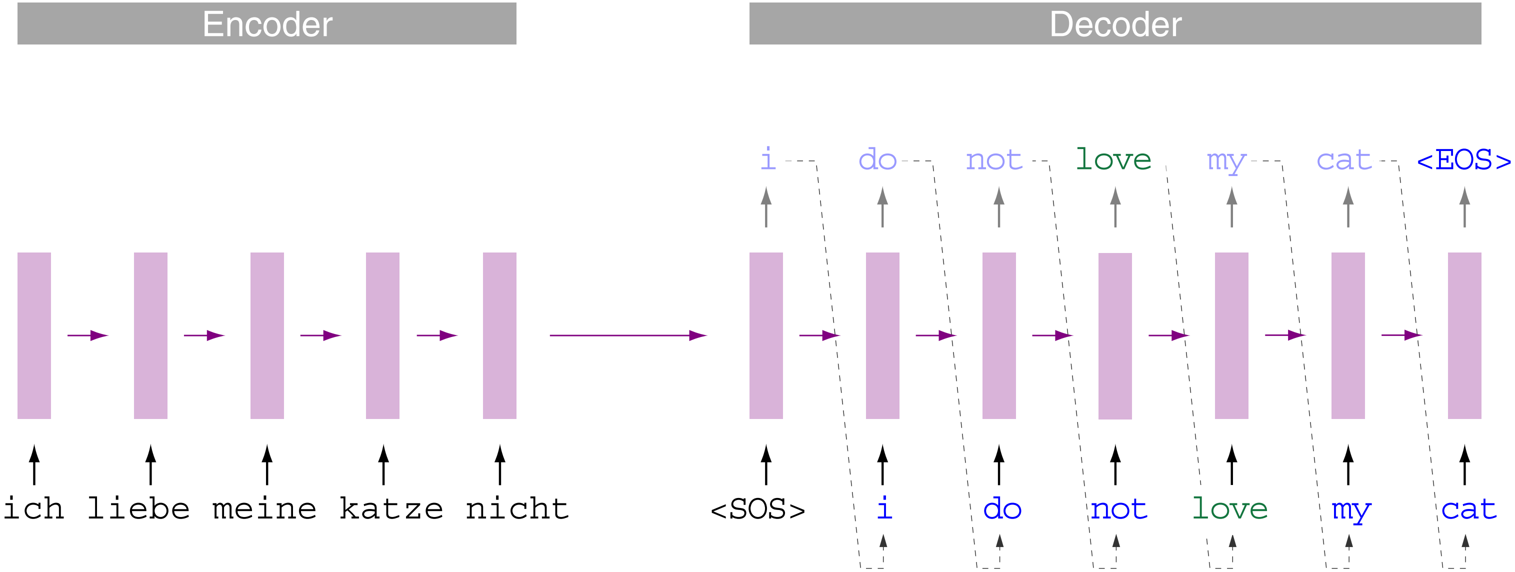
Selected basis technology

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.



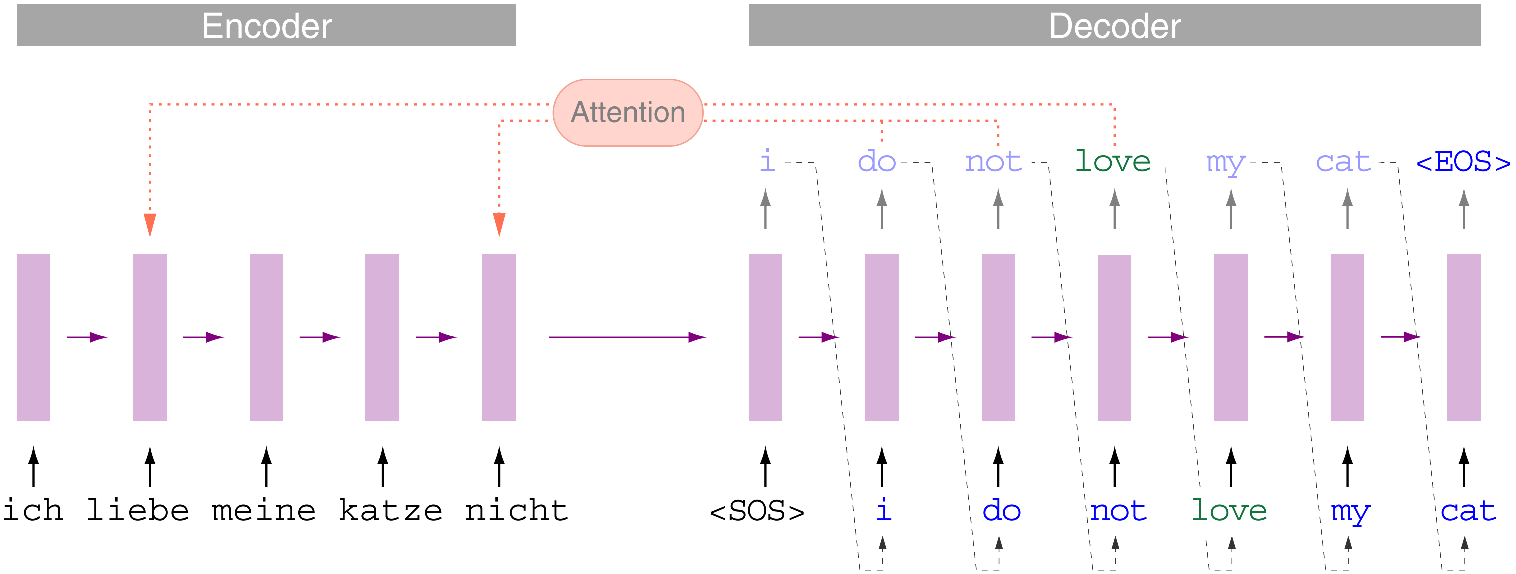
Selected basis technology

Neural language model (Bengio et al. 2000)

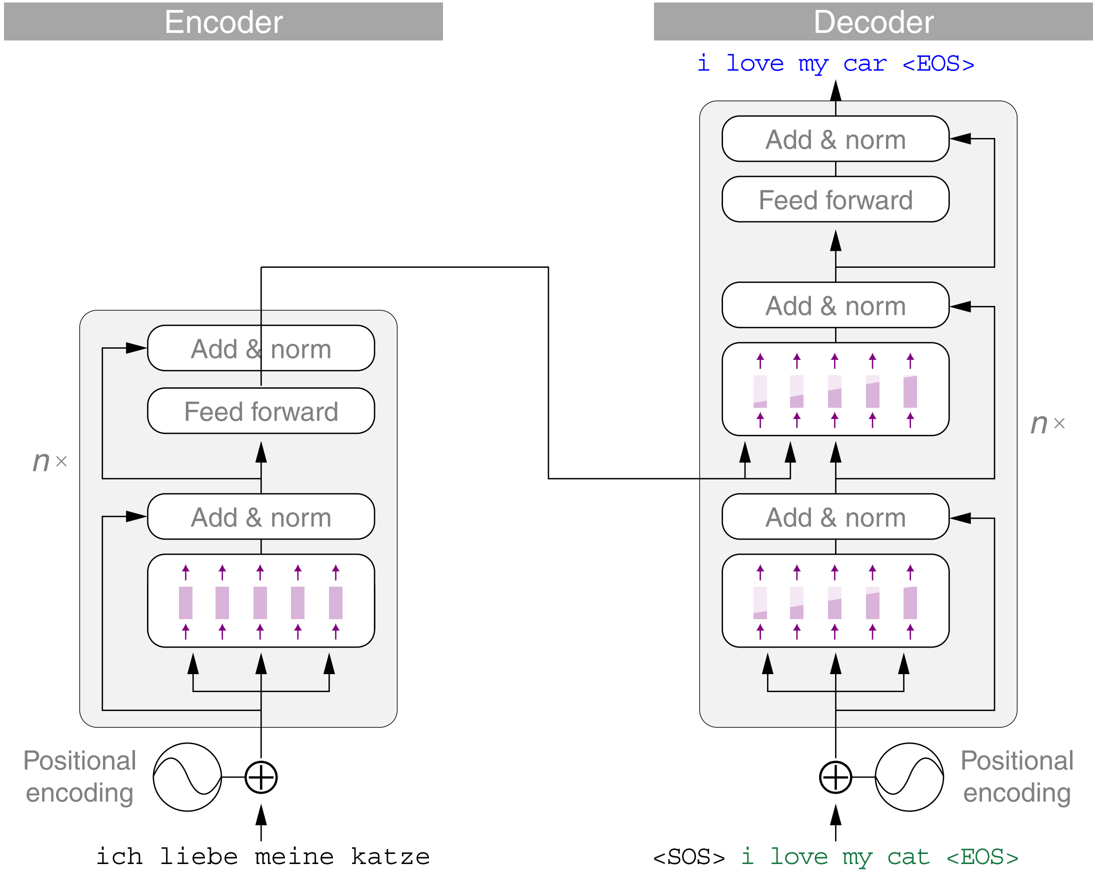
Recurrent neural language model with attention (Bahdanau et al. 2014)

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.



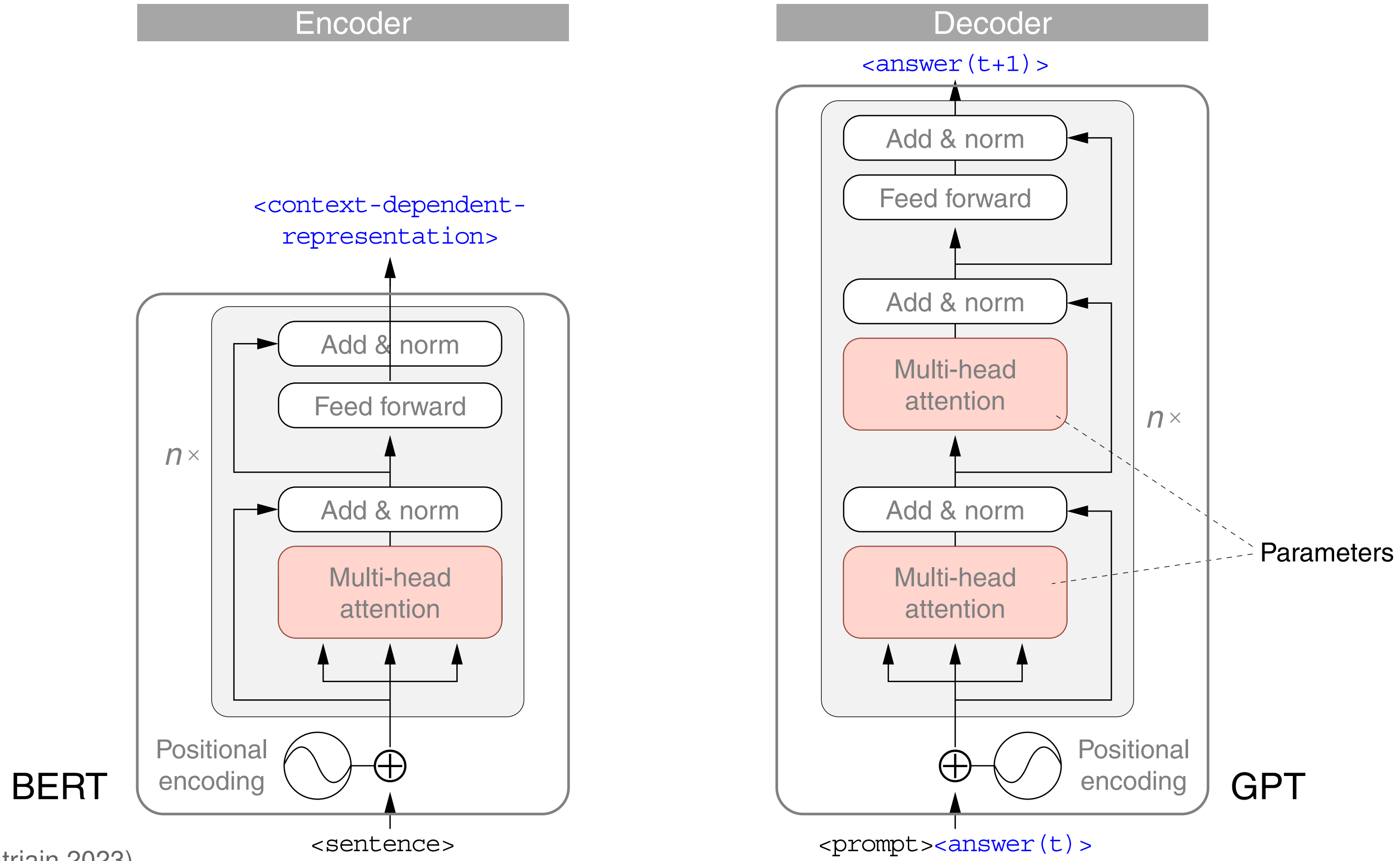
The Transformer (Vaswani et al., Google 2017)



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



Encoder

<context-dependent-representation>

Add & norm
Feed forward
Add & norm
Multi-head attention
$n \times$
Positional encoding
BERT

<sentence>

Decoder

<answer(t+1)>

Add & norm
Feed forward
Add & norm
Multi-head attention
Add & norm
Multi-head attention
$n \times$
Parameters
Positional encoding
GPT

<prompt><answer(t)>

Transformer models catalog (Amatriain 2023)

Selected basis technology

InstructGPT (Ouyang et al., OpenAI 2022)

RLHF (Christiano et al., OpenAI, Google 2017)

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

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.
- ≈ 34 days on 1,024 x A100 GPUs.
- $4.6M costs a single training run.

GPT-3 [Jun. 2020]

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]

Selected basis technology

InstructGPT (Ouyang et al., OpenAI 2022)
RLHF (Christiano et al., OpenAI, Google 2017)

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

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]

Agenda

① Background on Large Language Models and Transformers

② Who is the Author? Generative LLM Authorship Verification

③ Turing X (interactive)

④ The Infobot Project – An LLM-based Teaching Prototype for Lectures

⑤ Watermarking Large Language Models

The Daily Telegraph

BRITAIN'S BEST-SELLING QUALITY DAILY

Dialogue and co-operation are the only way to cyber security

[redacted]

As one of the greatest inventions in the 20th century, the internet has brought profound changes to our way of thinking, working and living. At the same time, it is prone to security risks and challenges. Wiretapping, attacks and terrorism in cyberspace have become global problems that call for global solutions. This means countries must work together instead of accusing one country for all the problems as some countries recently did against China, not to mention how

a lawless land. No country would tolerate fraud, cheating, stealing, terrorism or incitement of religious extremism.

The Chinese government has no part in stealing commercial secrets, nor do we in any way encourage or support any individual or company to do so. On the contrary, China has been opposing and cracking down on all forms of cyber theft all along.

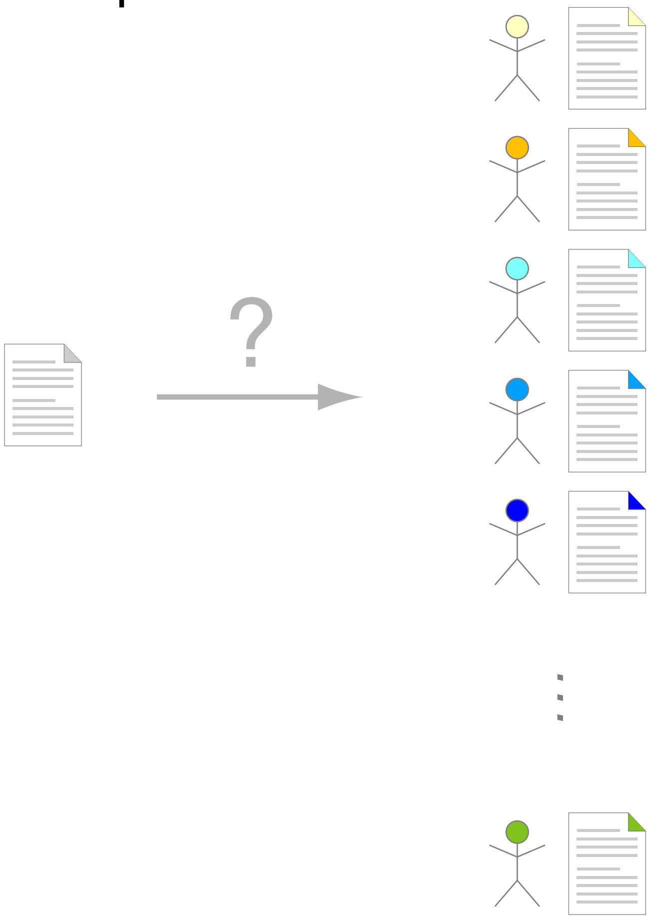
In recent years, China has strengthened rule of law in cyberspace and kept improving the relevant laws and regulations: the Cyber Security Law and The National Cyberspace Security Strategy were issued in 2016; The first internet court was established in Hangzhou in 2017, followed by the second and third in Beijing and

As a responsible big country, China has been actively pushing for bilateral and multilateral cooperation on cyber security, engaging with the US, the UK and the EU through dialogue mechanisms, and sharing China's wisdom at the UN and the G20 on improving international cooperation in cyberspace. Moreover, China has hosted five sessions of the World Internet Conference since 2014 to promote international cooperation on cyber security and cyber governance.

All these show that the accusations against China on cyber security are unfair, groundless and the opposite of the fact. People of the world need not be reminded who has conducted massive cyber wiretapping against foreign governments – even allies, who has engaged in organised cyber theft

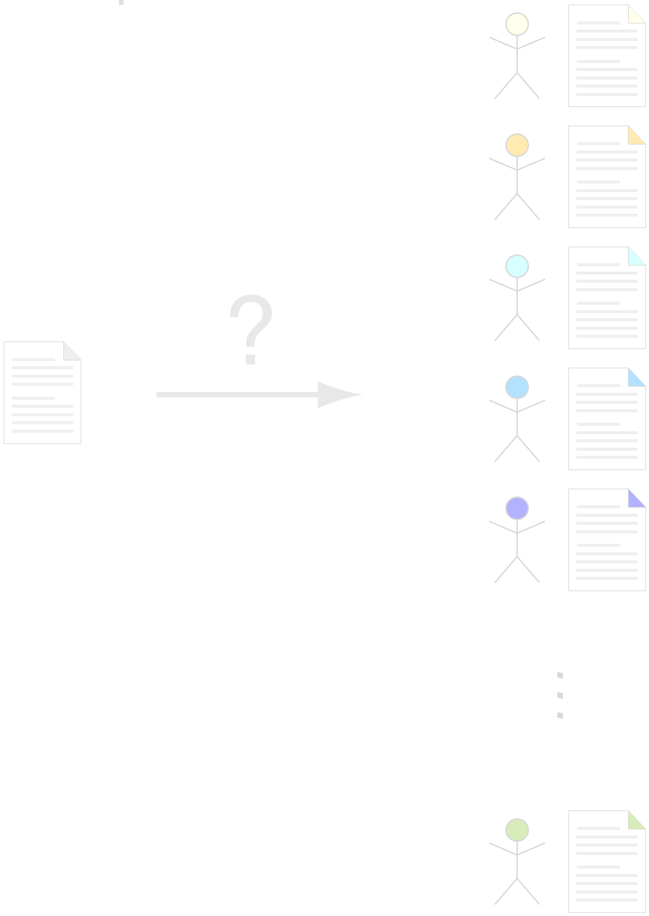
Graeme Hirst

Authorship Attribution



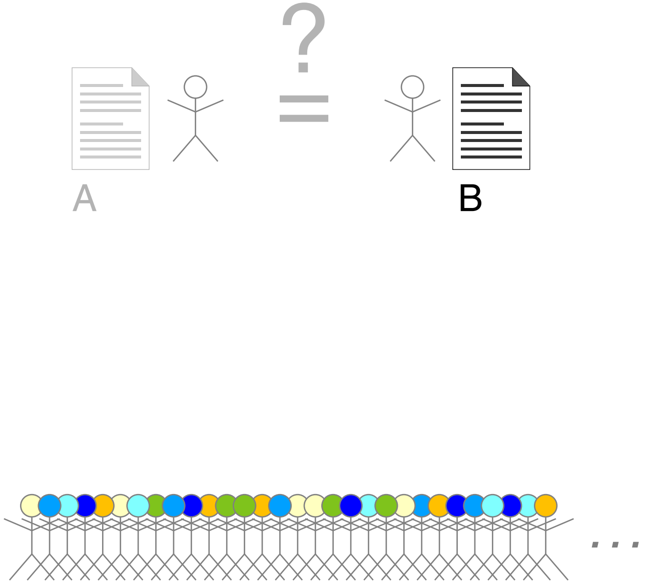
To which author does a text belong?

Authorship Attribution



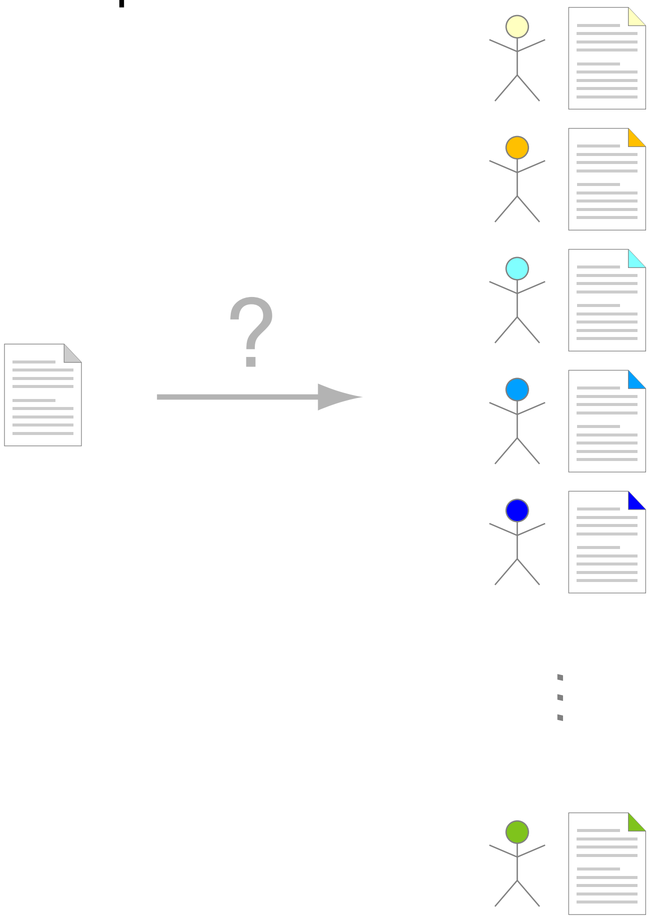
To which author does a text belong?

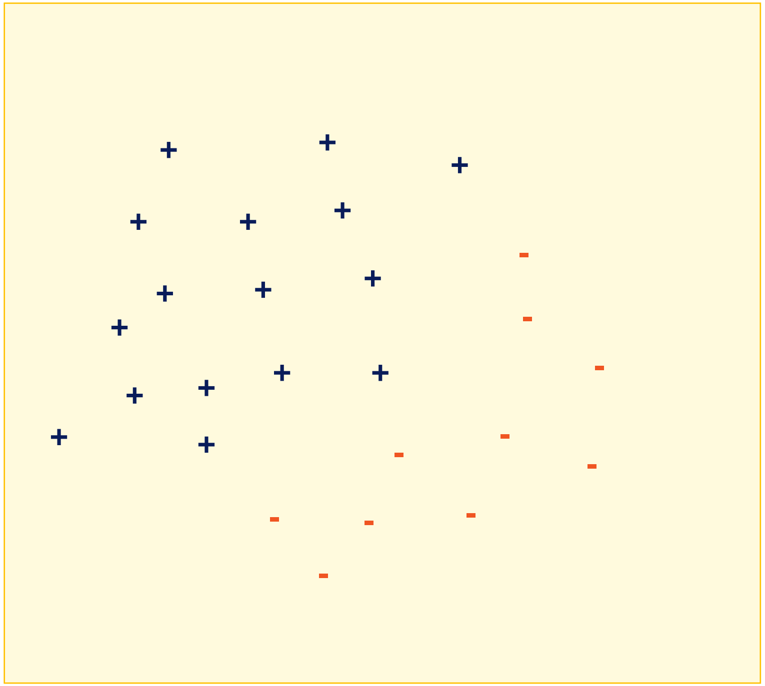
Authorship Verification



Originate two texts from the same author?

Authorship Attribution

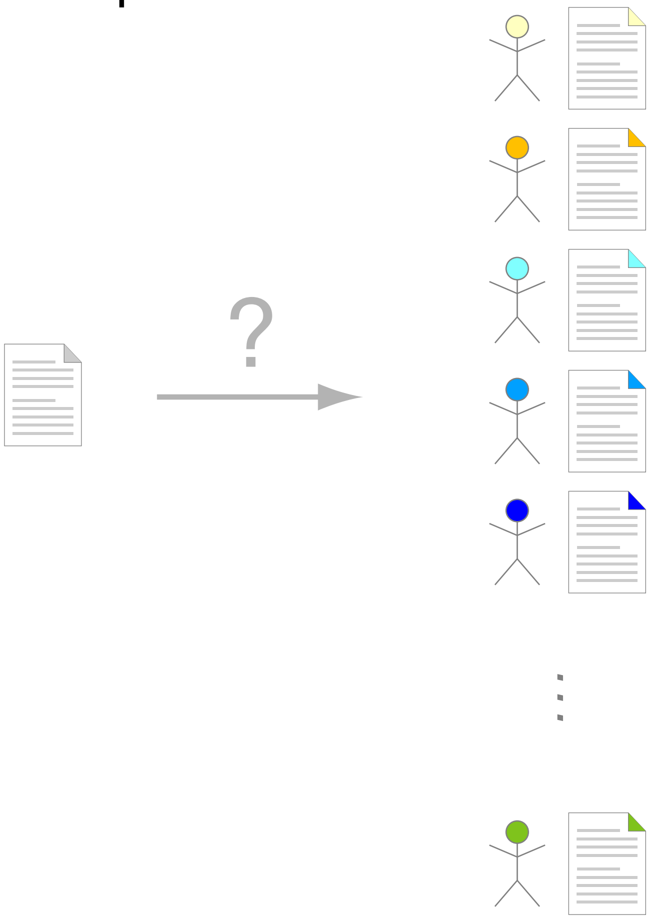


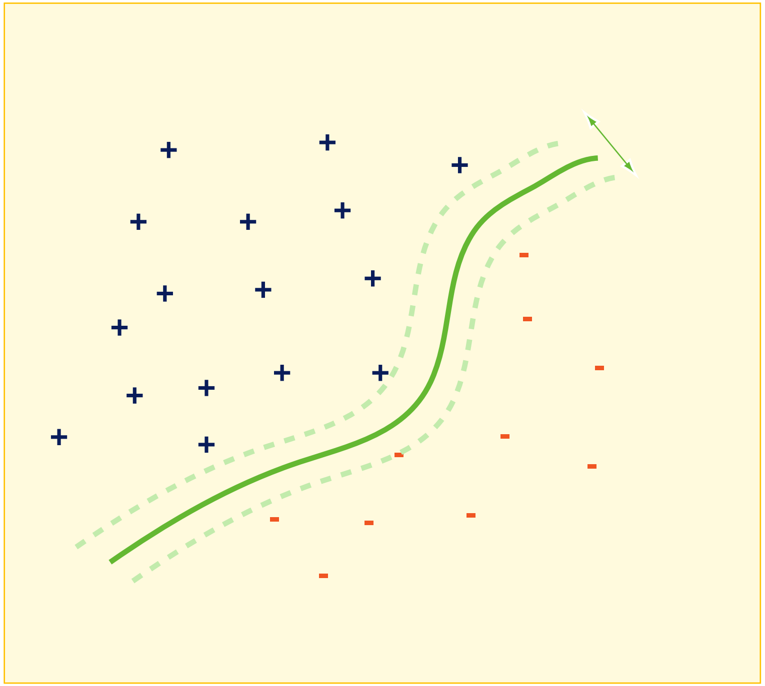


Discrimination-based classification.

To which author does a text belong?

Authorship Attribution

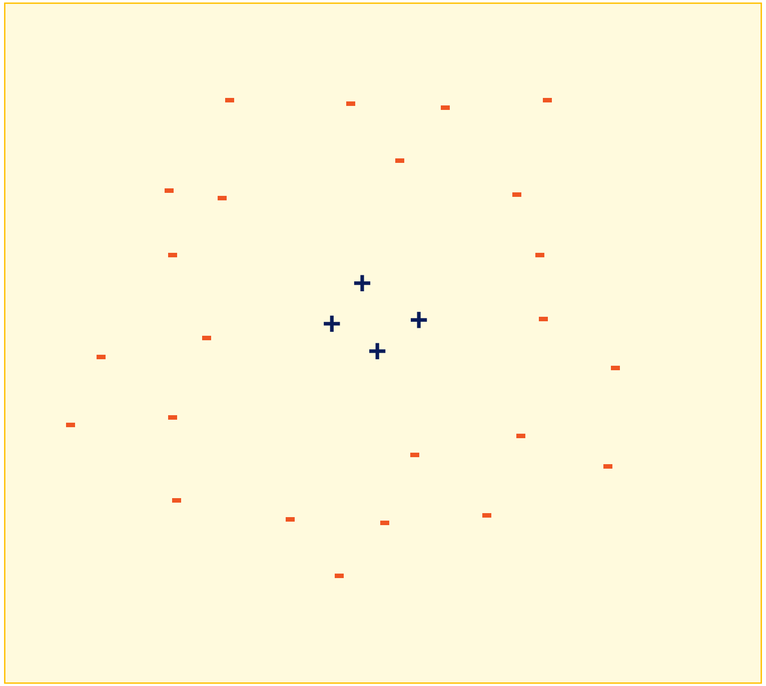




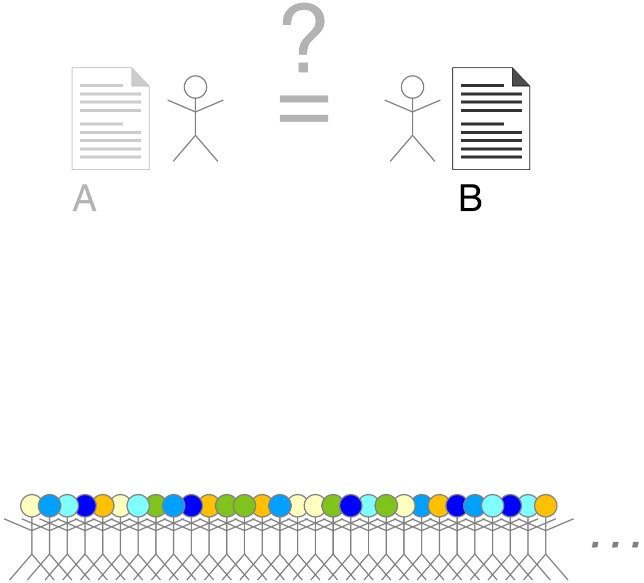
Discrimination-based classification.

To which author does a text belong?

Authorship Verification

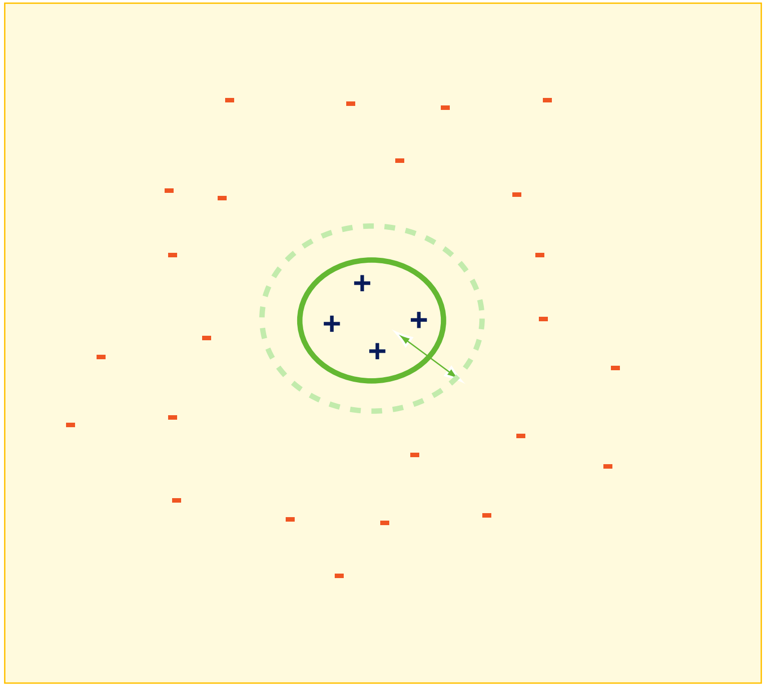


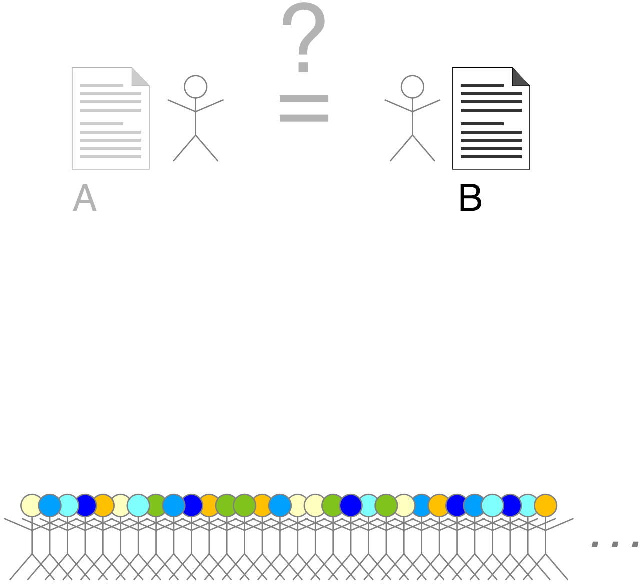
One-class classification.



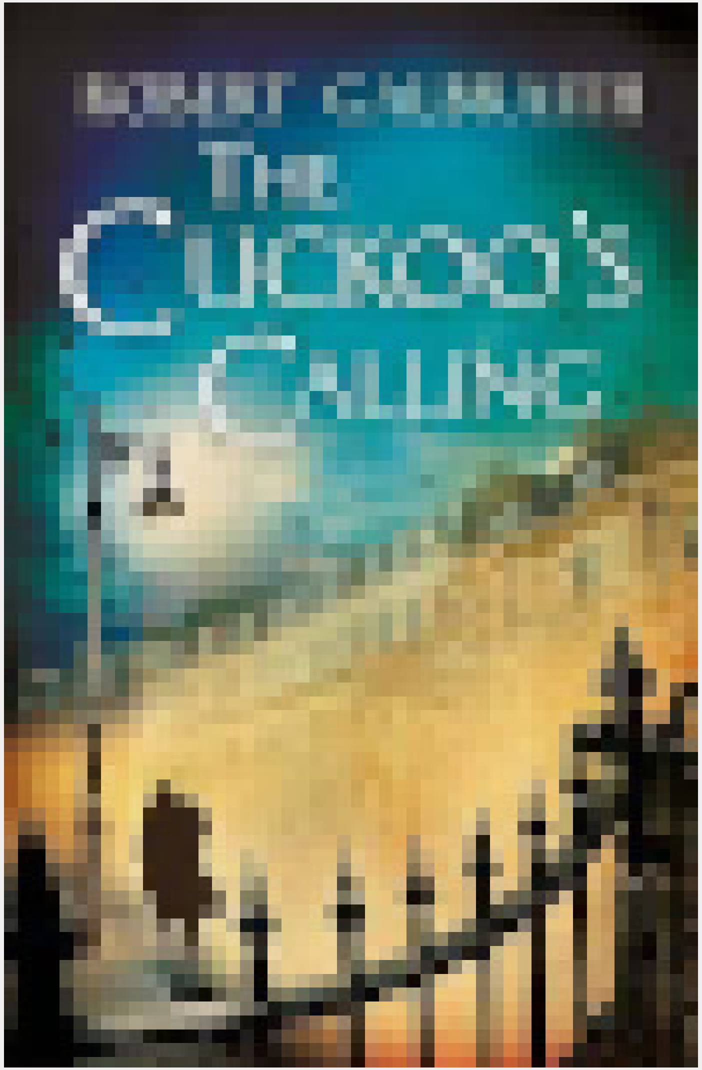
Originate two texts from the same author?

Authorship Verification

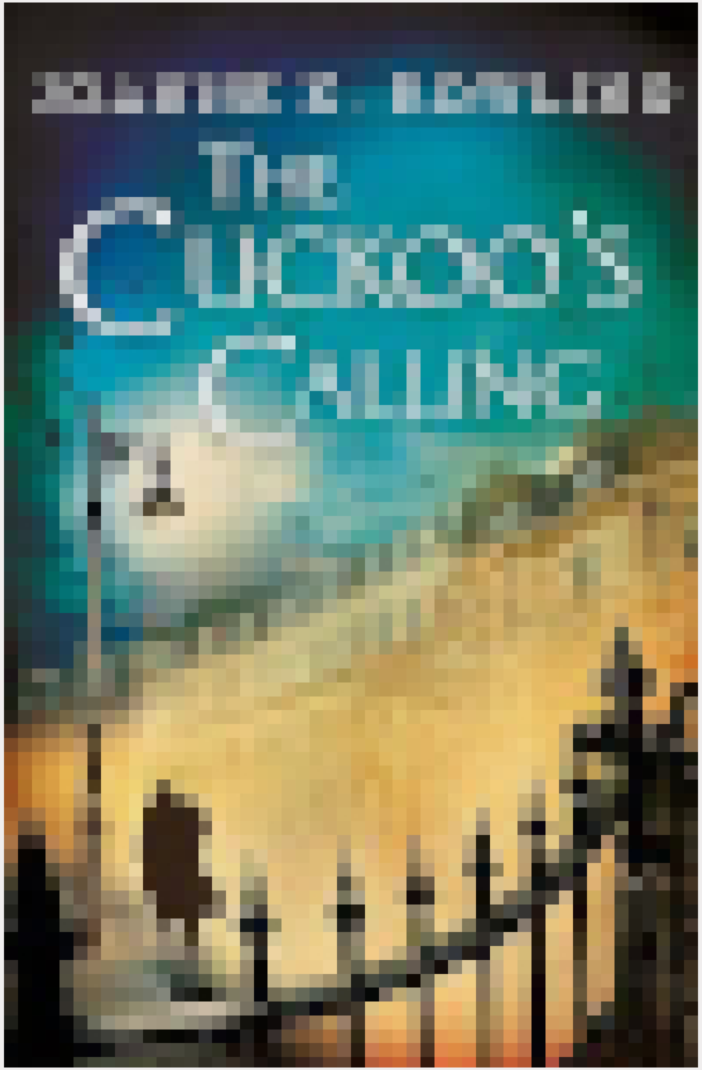

One-class classification.



Originate two texts from the same author?



?
=



2013, Patrick Juola

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

`The migrants who sailed with Gilbert were better fitted`
`for a crusade than a colony, and, disappointed at not`
`at once finding mines of gold and silver, many deserted ...`

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

T**he** migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

Th**e m**igrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

The **mi**grants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

```
The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...
```

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

The migrants who sailed with Gilbert were better fitted
for a crusade than a colon**y,** and, disappointed at not
at once finding mines of gold and silver, many deserted ...

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

The migrants who sailed with Gilbert were better fitted
for a crusade than a colony**, a**nd, disappointed at not
at once finding mines of gold and silver, many deserted ...

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

```
The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...
```

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

`The migrants who sailed with Gilbert were better fitted for a crusade than a colony,` **`and`**`, disappointed at not at once finding mines of gold and silver, many deserted ...`

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, **and**, disappointed at not
at once finding mines of gold and silver, many deserted ...

Author A

beautiful christmas you know jesus our saviour was born here below, patiently stooping to hunger and pain, so he might save us, his lost ones, from shame; now if we love him, he bids us to feed all his poor brothers and sisters who need. blessed old nick! i was sure if . . .

Trigram Freq.

Author B

come and see zip, the foremost of freaks! come and see palestine’s sinister sheiks! eager equestriennes, each unexcelled, most mammoth menagerie ever beheld, the giant, the fat girl, the lion-faced man, aerial artists from far-off japan, audacious acrobats shot from a gun, don’t . . .

Trigram Freq.

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

`The migrants who sailed with Gilbert were better fitted`
`for a crusade than a colony,` **`and`**`, disappointed at not`
`at once finding mines of gold and silver, many deserted ...`

Author A

beautiful christmas you know jesus our saviour was born here below, patiently stooping to hunger and pain, so he might save us, his lost ones, from_shame; now if we love him, he bids us to_feed all his poor brothers and sisters who need. blessed old nick! i was sure if . . .

Trigram	Freq.
`and`	4
`to_`	3
`the`	1
`our`	5
`_sh`	1
. . .	

Author B

come and see zip, the foremost of freaks! come and see palestine’s sinister_sheiks! eager equestriennes, each unexcelled, most mammoth menagerie ever beheld, the giant, the fat girl, the lion-faced man, aerial artists from far-off japan, audacious acrobats shot from a gun, don’t . . .

Trigram	Freq.
`and`	2
`to_`	1
`the`	4
`our`	1
`_sh`	2
. . .	

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

`The migrants who sailed with Gilbert were better fitted`
`for a crusade than a colony,` **`and`**`, disappointed at not`
`at once finding mines of gold and silver, many deserted ...`

Author A

beautiful christmas you know jesus our saviour was born here below, patiently stooping to hunger and pain, so he might save us, his lost ones, from_shame; now if we love him, he bids us to_feed all his poor brothers and sisters who need. blessed old nick! i was sure if . . .

Trigram	Freq.
`and`	4
`to_`	3
`the`	1
`our`	5
`_sh`	1
...	

Author B

come and see zip, the foremost of freaks! come and see palestine’s sinister_sheiks! eager equestriennes, each unexcelled, most mammoth menagerie ever beheld, the giant, the fat girl, the lion-faced man, aerial artists from far-off japan, audacious acrobats shot from a gun, don’t . . .

Trigram	Freq.
`and`	2
`to_`	1
`the`	4
`our`	1
`_sh`	2
...	

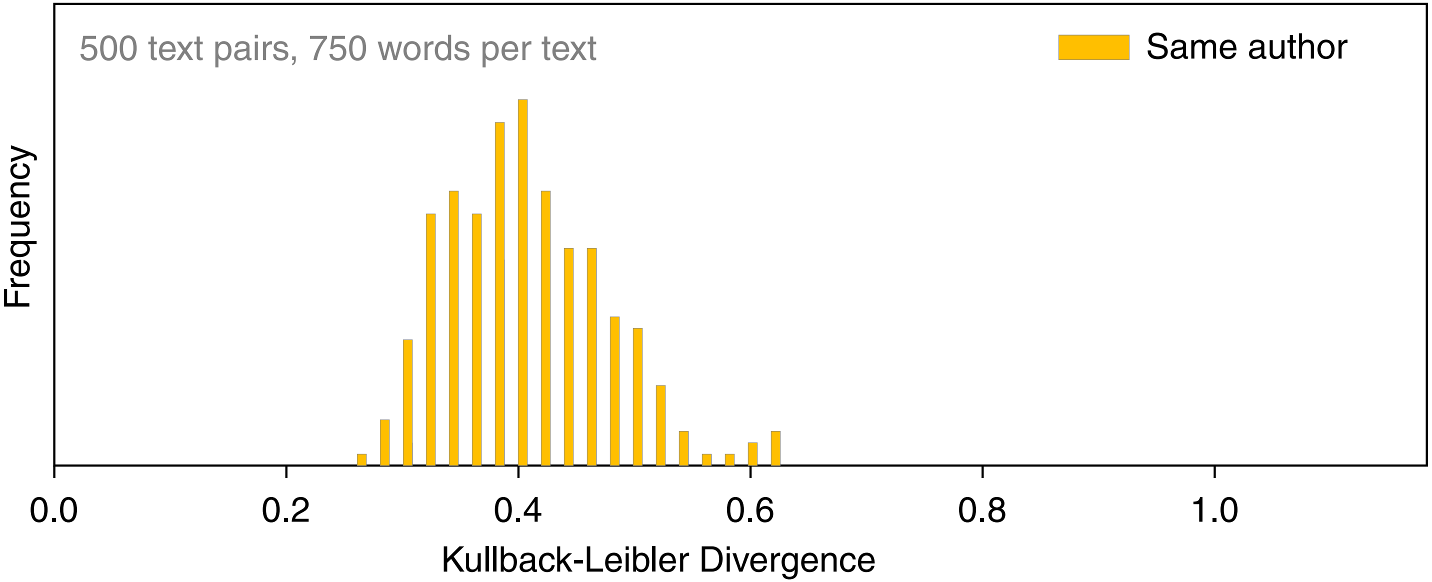
Kullback-Leibler Divergence:

$$\mathrm{KLD}(P \mid Q) = \sum_{i \in \text{trigrams}} P[i] \log \frac{P[i]}{Q[i]}$$

Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

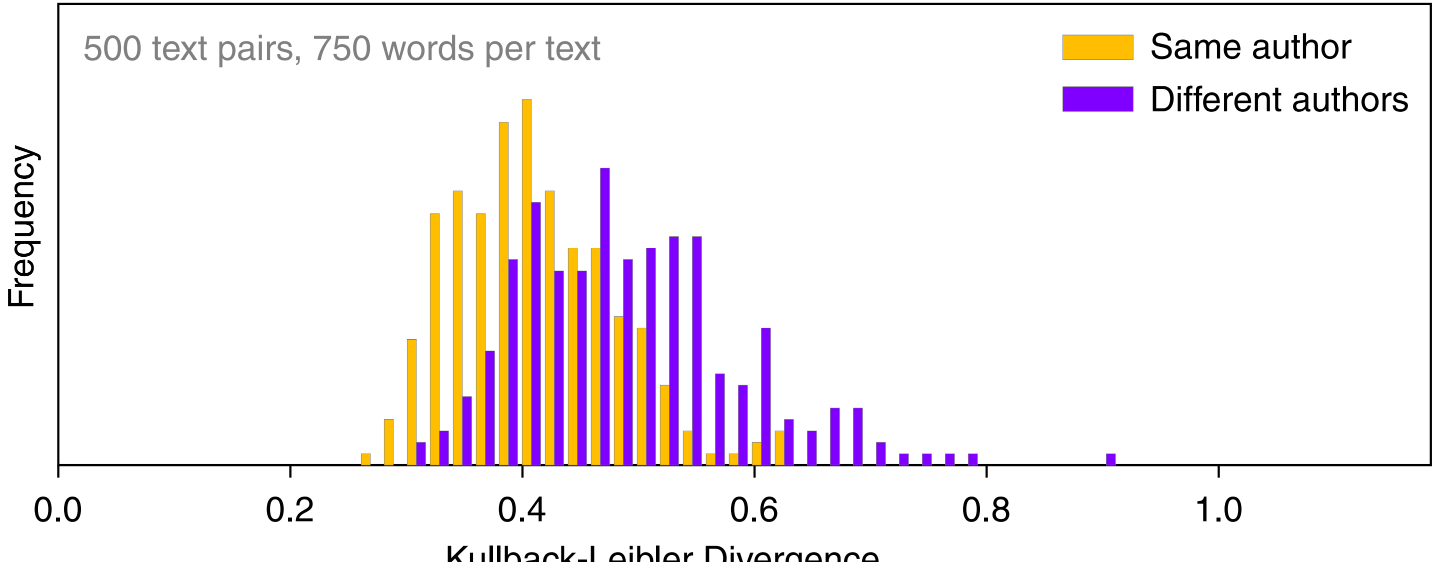
`The migrants who sailed with Gilbert were better fitted`
`for a crusade than a colony, and, disappointed at not`
`at once finding mines of gold and silver, many deserted ...`



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

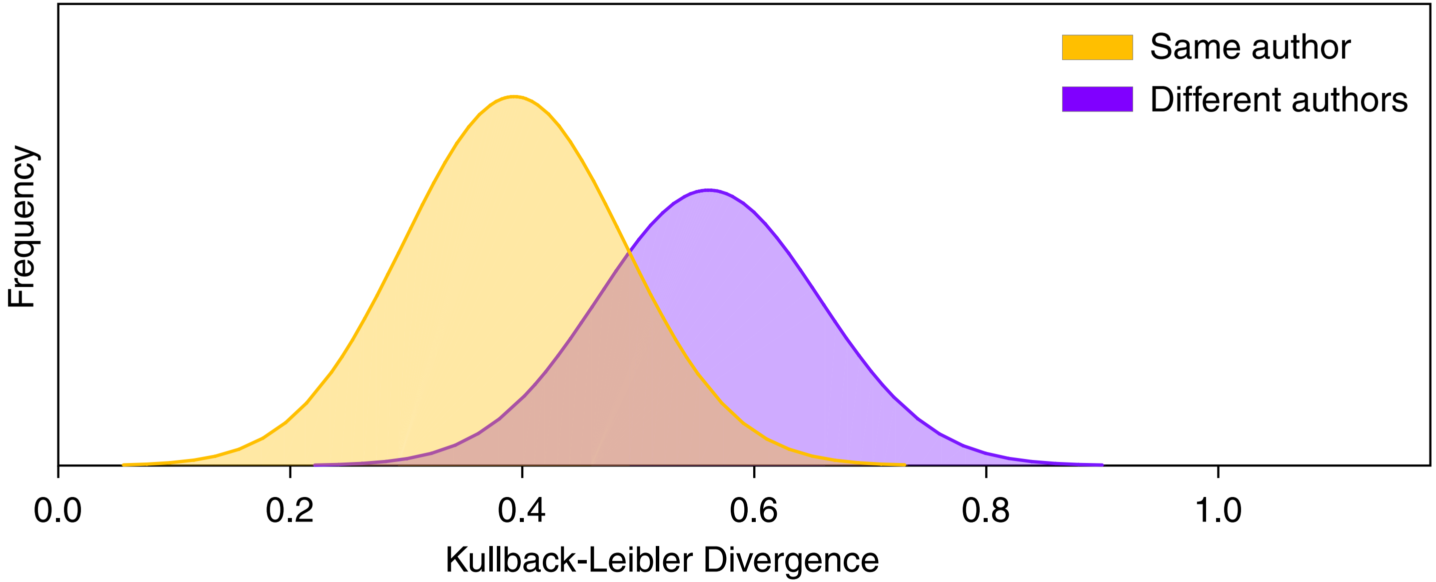
```
The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...
```



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

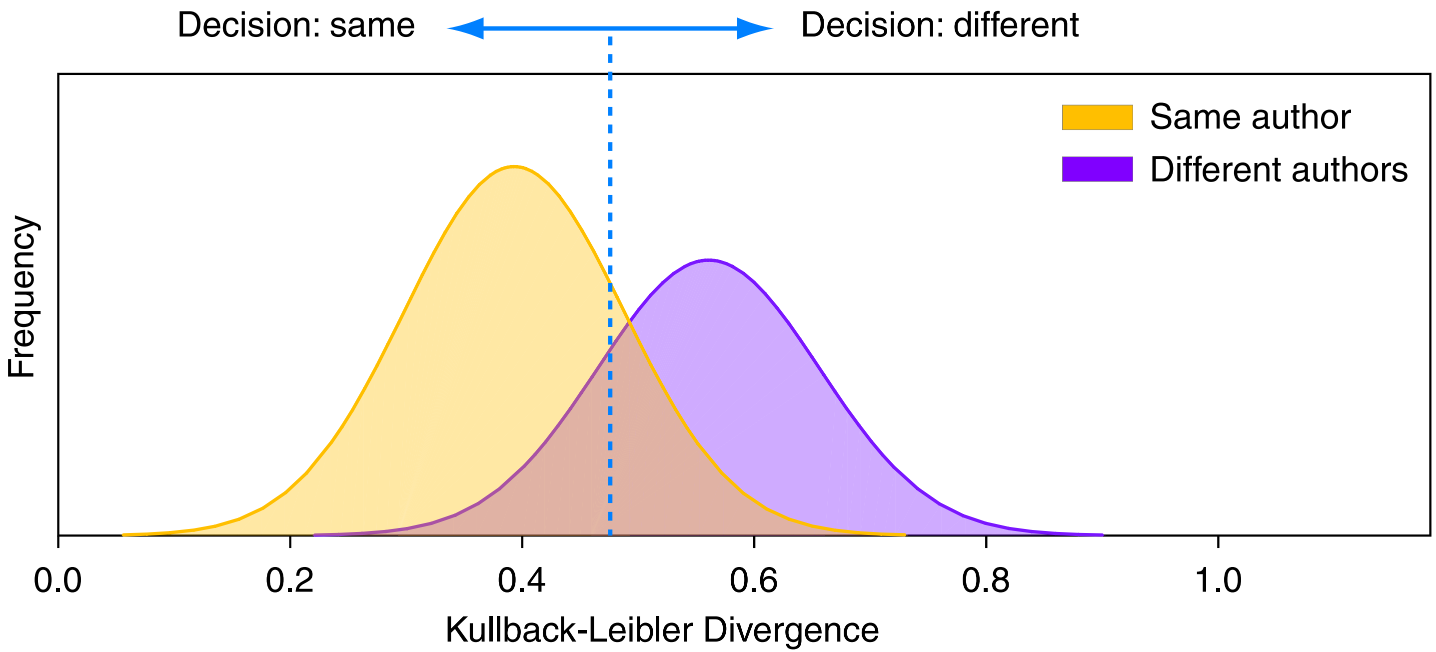
```
The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...
```



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

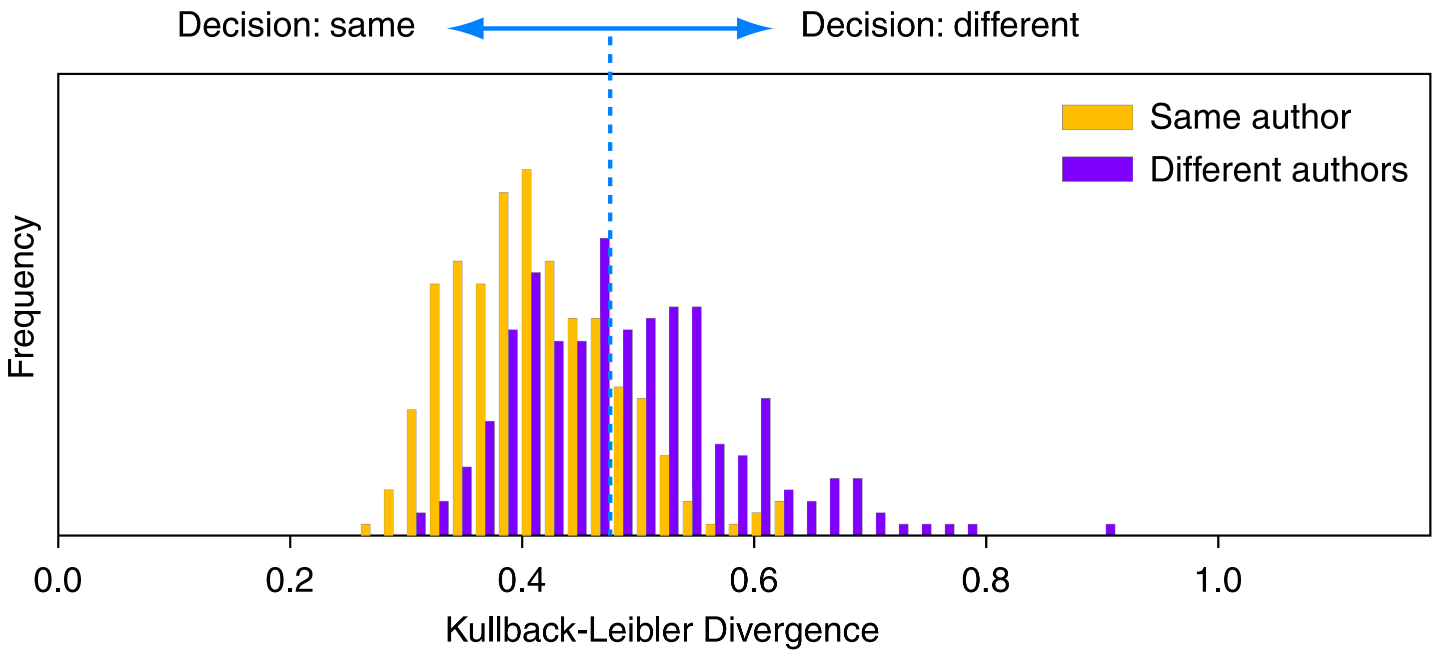
```
The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...
```



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

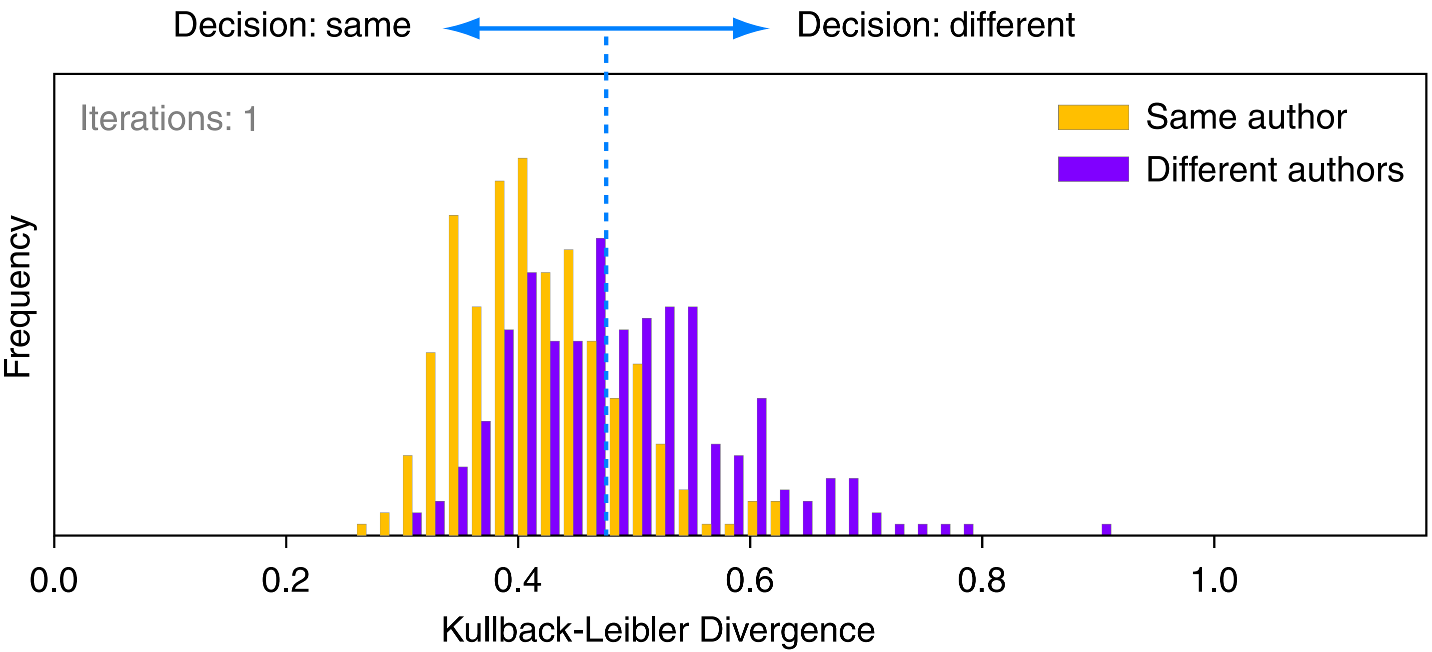
```
The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...
```



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

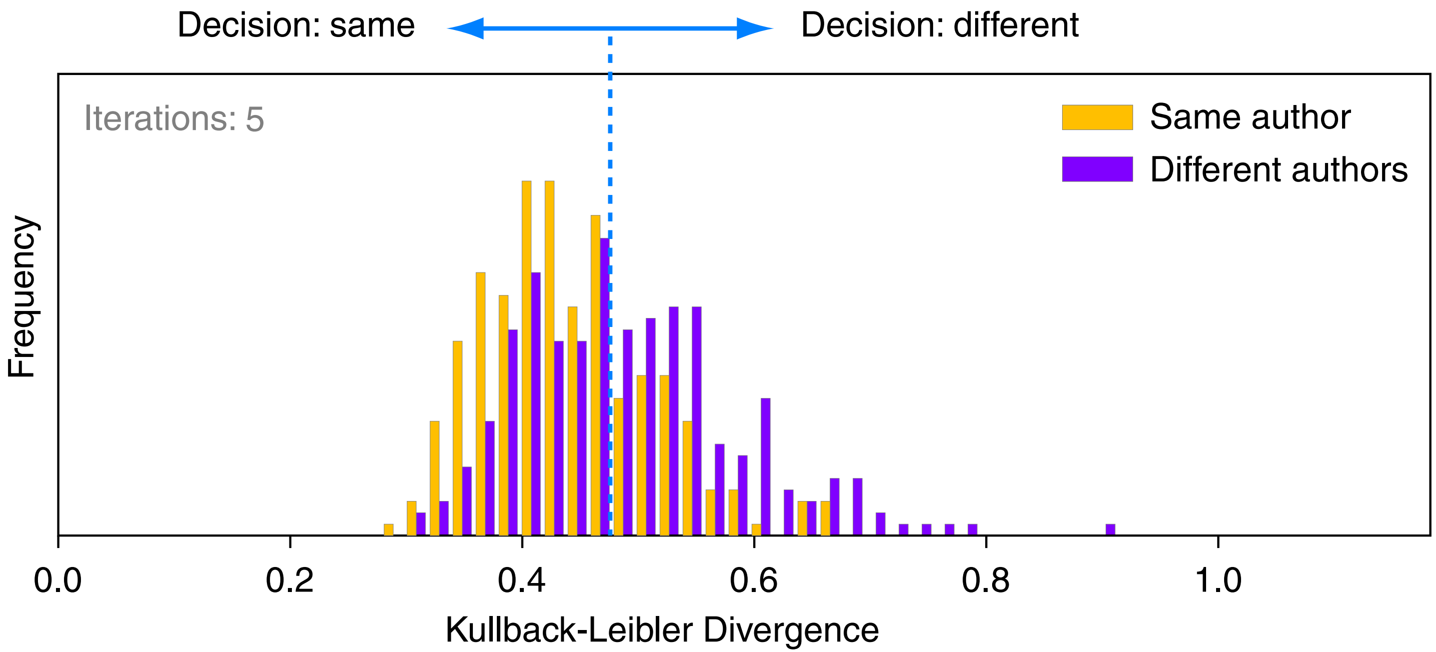
```
The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...
```



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

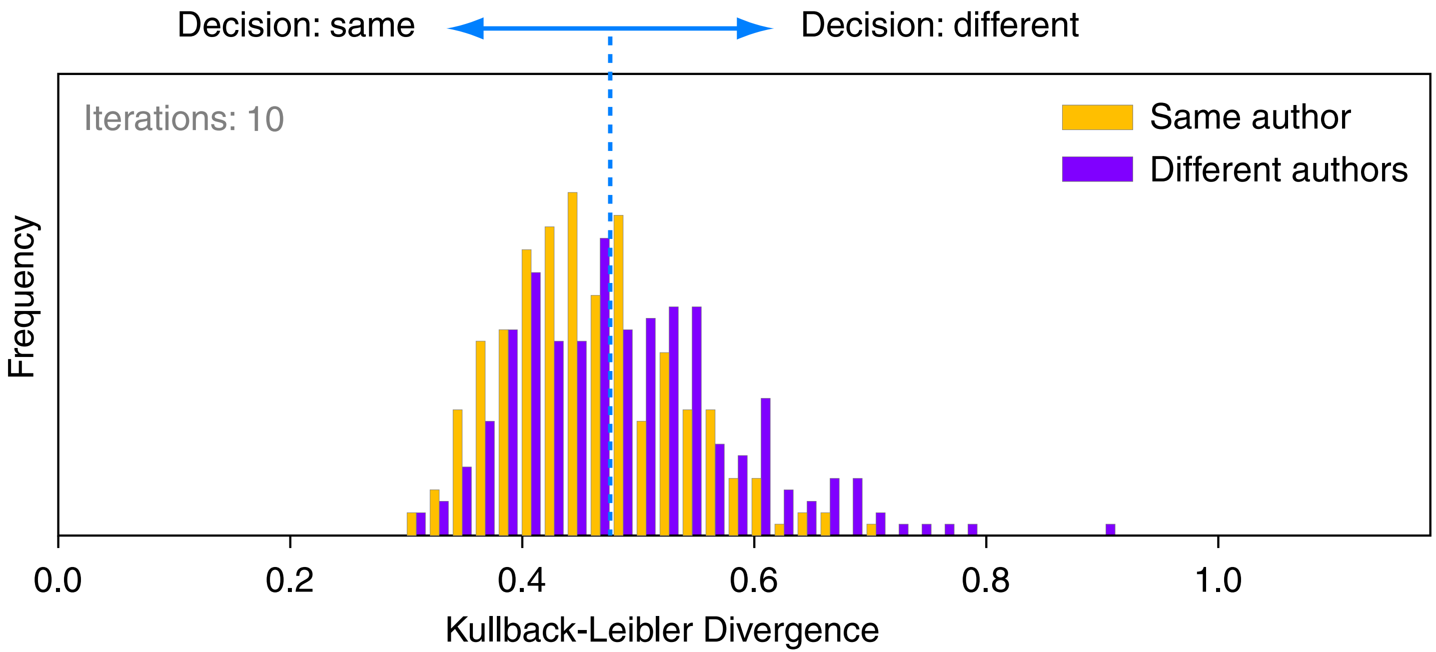
`The migrants who sailed with Gilbert were better fitted for a crusade than a colony, and, disappointed at not at once finding mines of gold and silver, many deserted ...`



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

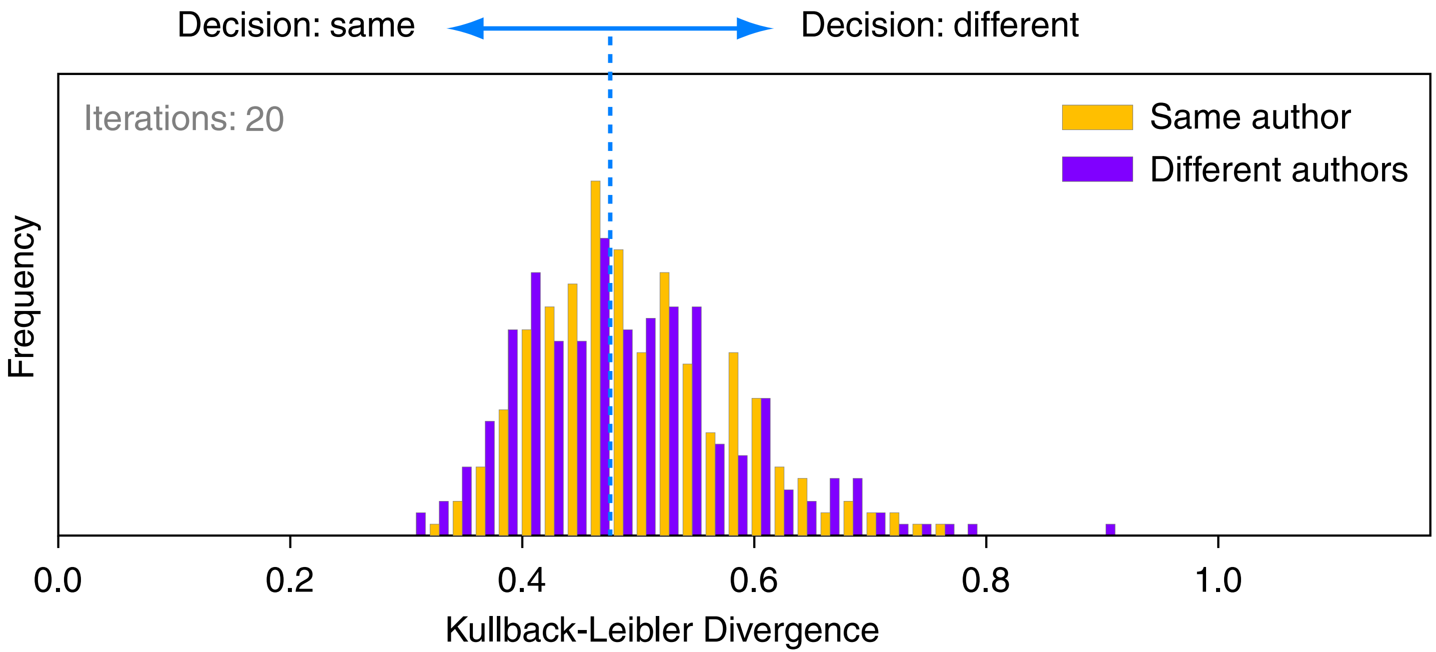
`The migrants who sailed with Gilbert were better fitted`
`for a crusade than a colony, and, disappointed at not`
`at once finding mines of gold and silver, many deserted ...`



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

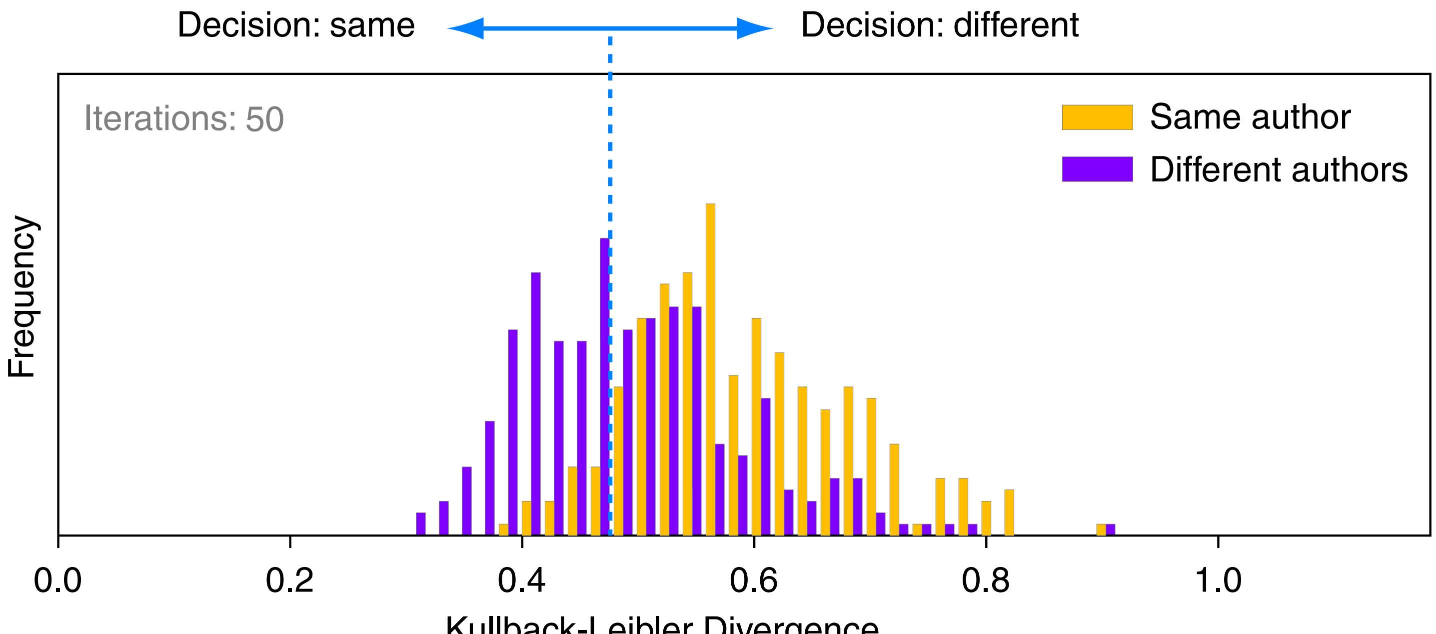
`The migrants who sailed with Gilbert were better fitted`
`for a crusade than a colony, and, disappointed at not`
`at once finding mines of gold and silver, many deserted ...`



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

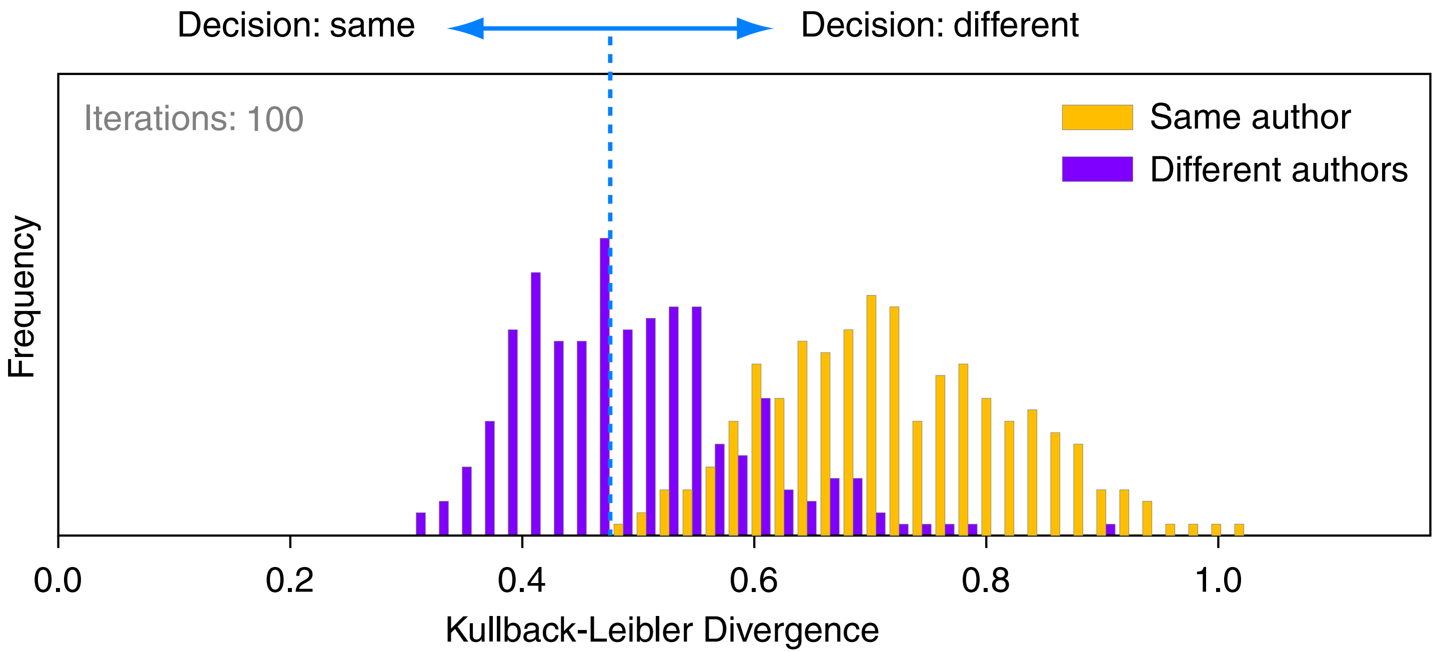
```
The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...
```



Authorship Analytics

Char-trigrams → sliding window with $n = 3$:

The migrants who sailed with Gilbert were better fitted
for a crusade than a colony, and, disappointed at not
at once finding mines of gold and silver, many deserted ...



PAN 2024 / 25

Lab on Digital Text Forensics and Stylometry

Janek Bevendorff

Jussi Karlgren

The Voight-Kampff * LLM Detection Task

[pan.webis.de]

---

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

Generative LLM Authorship Verification

Given two texts, one written by a human, the other by an LLM,
decide which text was written by whom.

Generative LLM Authorship Verification

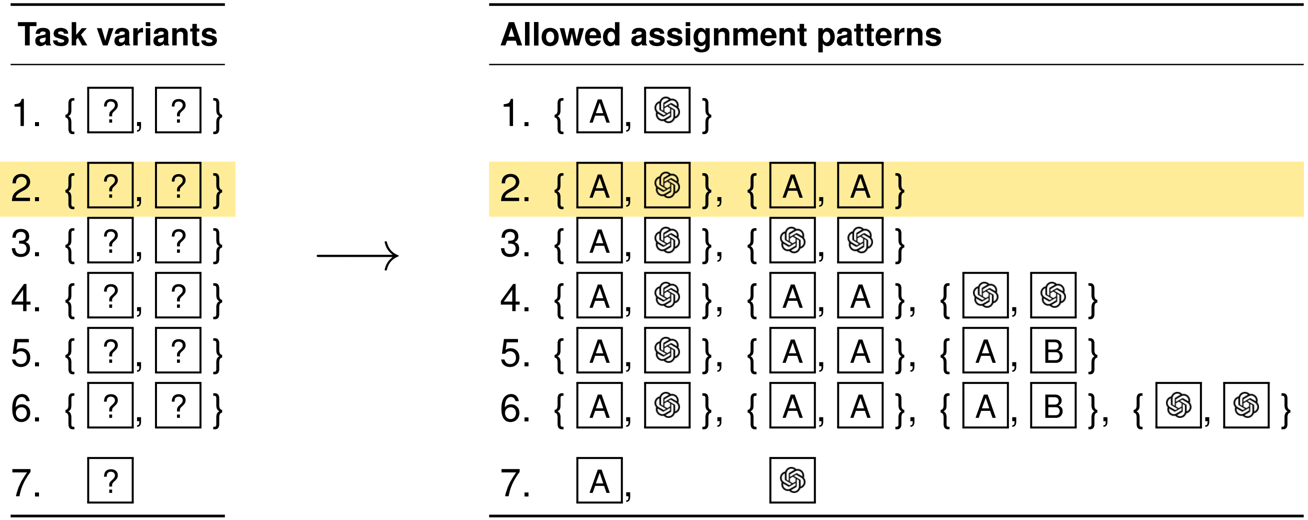
Given two texts, one written by a human, the other by an LLM, decide which text was written by whom.

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.

Generative LLM Authorship Verification

Given two texts, one written by a human, the other by a human or an LLM, 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.

Generative LLM Authorship Verification

Given a (potentially obfuscated) text,
decide whether it was written by a human or an LLM.

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.

Generative LLM Authorship Verification (dataset creation)

Human Texts : Curation of corpora from different genres.

(a) 7,300 19th-century novels (500–700 words).
Scraped from Project Gutenberg.

(b) 931 essays.
Brennan-Greenstadt (Brennan et. al, 2012) and Riddell-Juola (Wang et al., 2021) corpora.

(c) 870 news articles from 2021.
Crawled from Google News (also used at PAN’24).

(d) 22 texts of mixed genres.
ELOQUENT dataset (only for test).

Generative LLM Authorship Verification (dataset creation)

Machine Texts: Reconstruction of human texts by 14 LLMs.*

1. Decompose human texts.
   - `"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', ..."`

2. Synthesize new texts.
   - `"You are an essay summarizer and a forensic writing style analyst ..."`
   - `"If the essay is argumentative, classify the author's stance ..."`
   - `"Use very short sentences."` `"Use passive voice a lot."`
   - `"Write like a 7-year-old."` `"Write in Yoda grammar."`

3. Test data variants to analyze selected robustness aspects.
   Unicode obfuscation, cropped text (35 words), cross-topic pairs, cross-language pairs.

4. The generated texts are cleaned manually of artifacts.

---

* 14 state-of-the-art LLMs, among others GPT-3.5, GPT-4o, GPT-4o-mini, Gemini, DeepSeek, Llama

Generative LLM Authorship Verification (baselines and submissions)

- ❑ 3 Baseline systems:
  - Binoculars [Hans et al., 2024]
  - PPMd Compression-based Cosine [Sculley and Brodly, 2006; Halvani et al., 2017]
  - SVM with TF-IDF features
- ❑ Evaluation measures:

  ROC-AUC, Brier, C@1, $F_{0.5u}$, F_1, Mean of all
- ❑ 24 Submissions (30 submissions in 2024)
- ❑ Top systems:
  - fine-tuned Qwen3 with training data obfuscation and model selection
  - ensemble of Qwen+ModernBERT; cumulative term-document correlation matrix
- ❑ Other approaches:

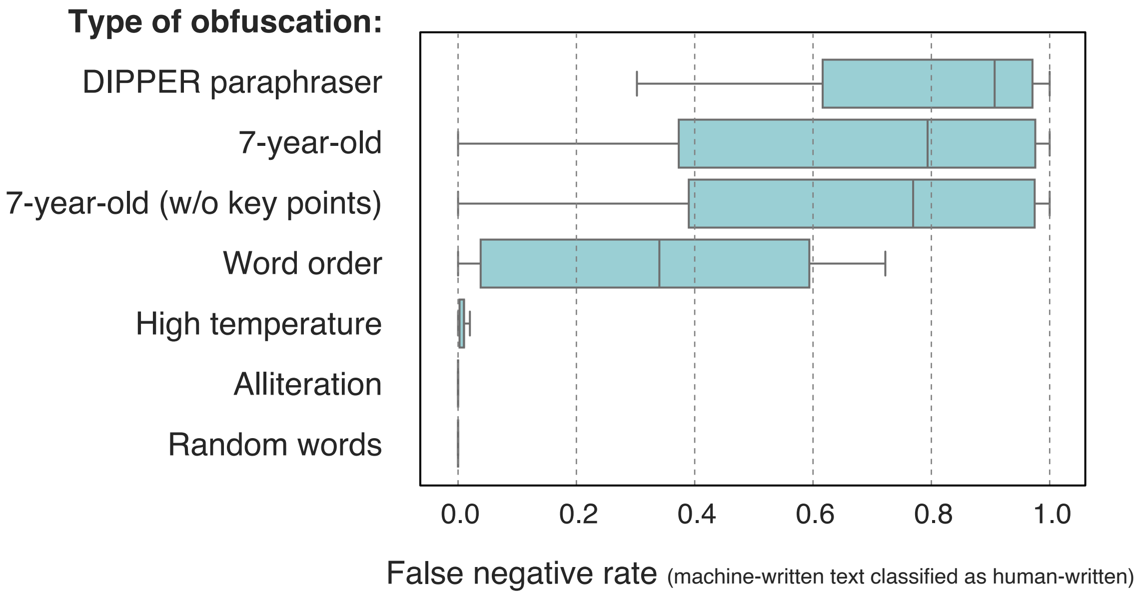
  LLM embeddings (13), stylometry (7), augmented data (6), ensembles (5), custom loss (5)

Generative LLM Authorship Verification (systems ranking)

	Team	ROC-AUC	C@1	F_1	Mean	FNR	FPR
1	Macko	**0.995**	**0.982**	**0.989**	**0.989**	**0.006**	**0.018**
2	Valdez-Valenzuela	0.939	0.897	0.926	0.929	0.020	0.107
3	Liu	0.962	0.889	0.923	0.928	0.005	0.120
4	Seeliger	0.912	0.896	0.930	0.925	0.082	0.103
5	Voznyuk	0.899	0.898	0.929	0.924	0.035	0.107
			⋮				
	Baseline TF-IDF SVM	0.963	0.897	0.904	0.922	0.106	0.093
			⋮				
17	Basani	0.904	0.843	0.894	0.891	0.084	0.160
			⋮				
	Baseline Binoculars	0.827	0.818	0.866	0.863	0.263	0.173
			⋮				
	Baseline PPMd CBC	0.644	0.759	0.817	0.790	0.797	0.137
24	Liang	0.734	0.694	0.752	0.751	0.157	0.298

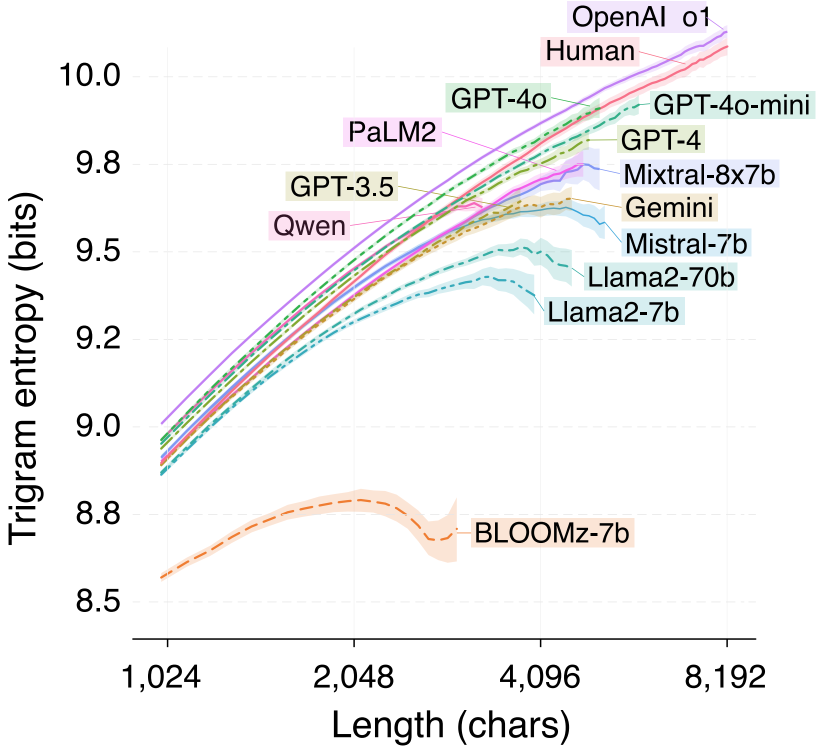
Generative LLM Authorship Verification (evaluation*)

Effect of data obfuscation on the top-10 systems:



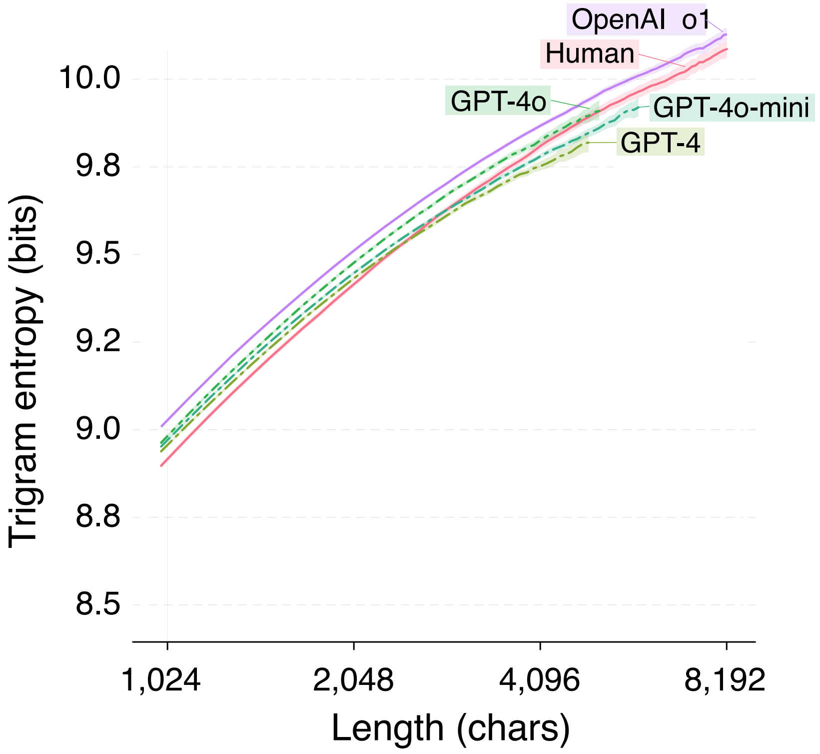
* J.Bevendorff et al. Overview of the 2nd 'Voight-Kampff' Generative AI Authorship Verification Task at PAN and ELOQUENT 2025. [CLEF 2025]

Generative LLM Authorship Verification (distinguishability in the future*)



* Bevendorff/Wiegmann/Richter/Potthast/Stein. The Two Paradigms of LLM Detection: Authorship Attribution vs. Verification. [ACL 2025]

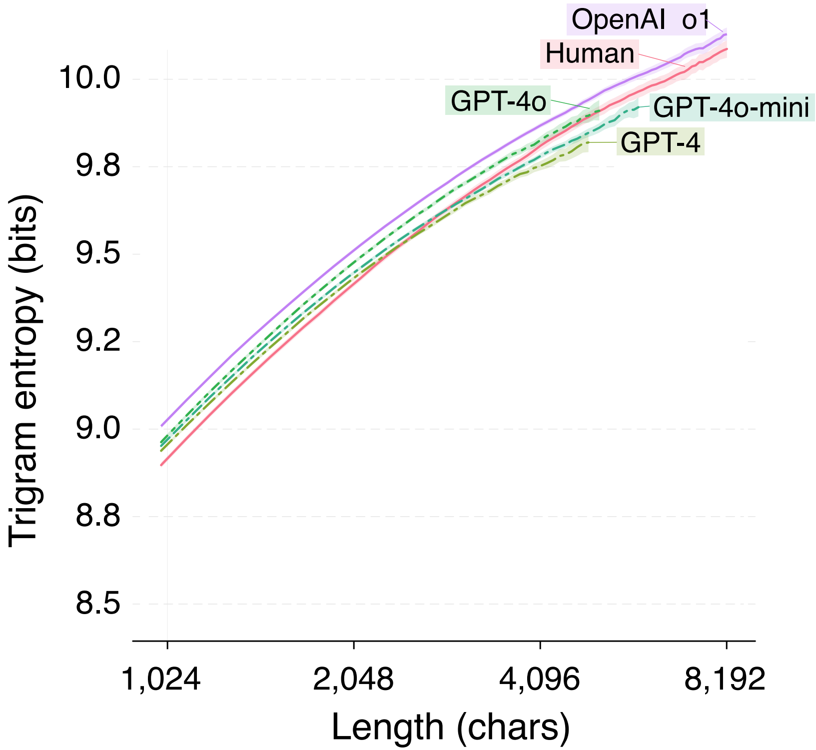
Generative LLM Authorship Verification (distinguishability in the future*)

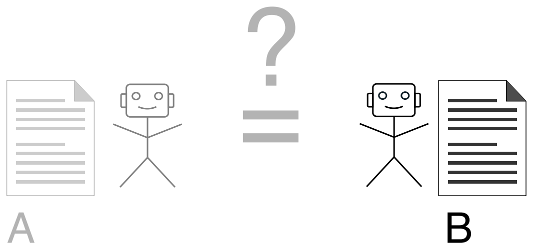


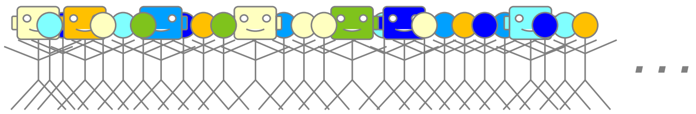
---

* Bevendorff/Wiegmann/Richter/Potthast/Stein. The Two Paradigms of LLM Detection: Authorship Attribution vs. Verification. [ACL 2025]

Generative LLM Authorship Verification (distinguishability in the future*)







Originate two texts from the same author, author $\in$ {human, LLM}?

---

* Bevendorff/Wiegmann/Richter/Potthast/Stein. The Two Paradigms of LLM Detection: Authorship Attribution vs. Verification. [ACL 2025]

Agenda

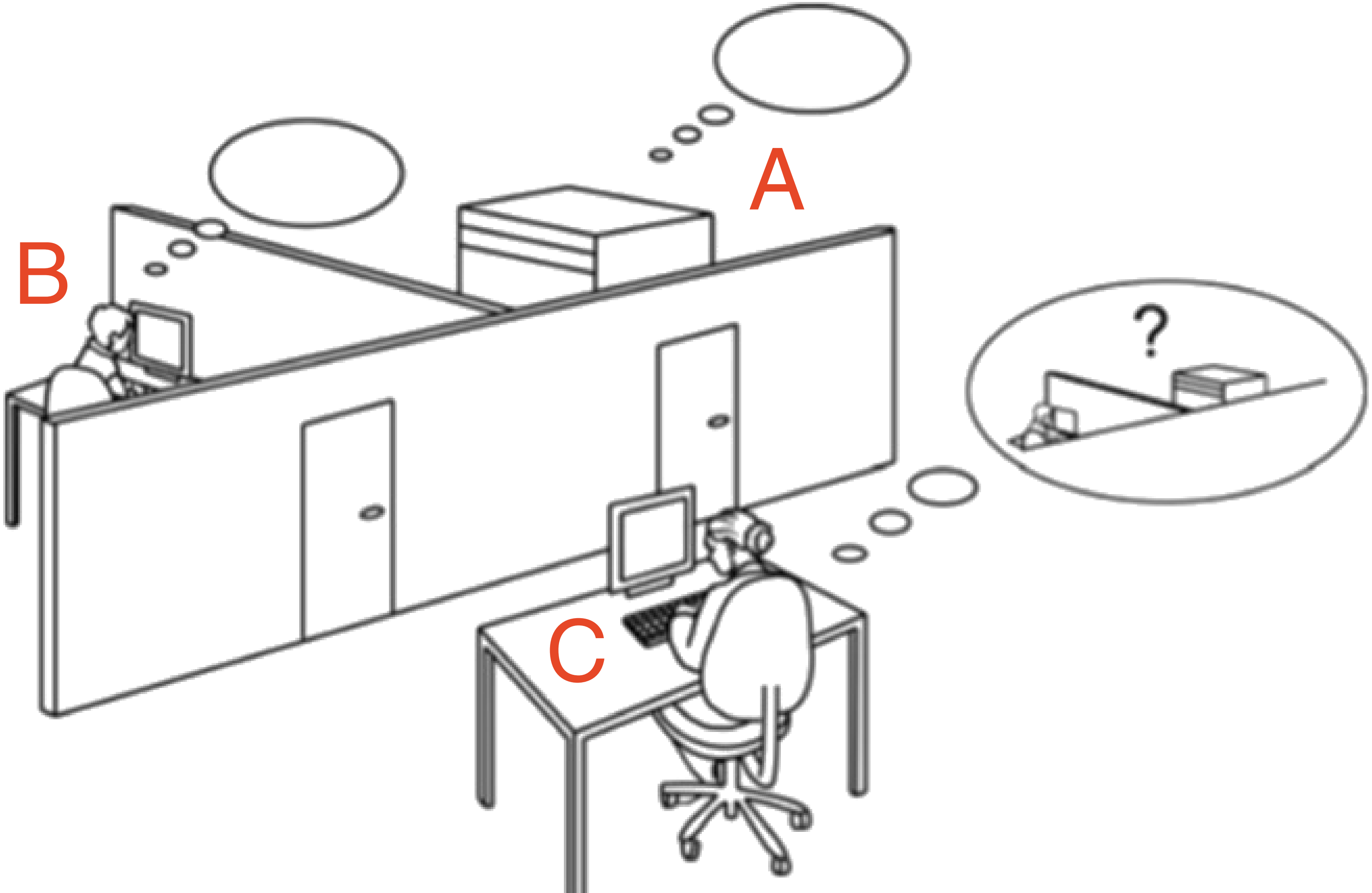
① Background on Large Language Models and Transformers

② Who is the Author? Generative LLM Authorship Verification

③ Turing X (interactive)

④ The Infobot Project – An LLM-based Teaching Prototype for Lectures

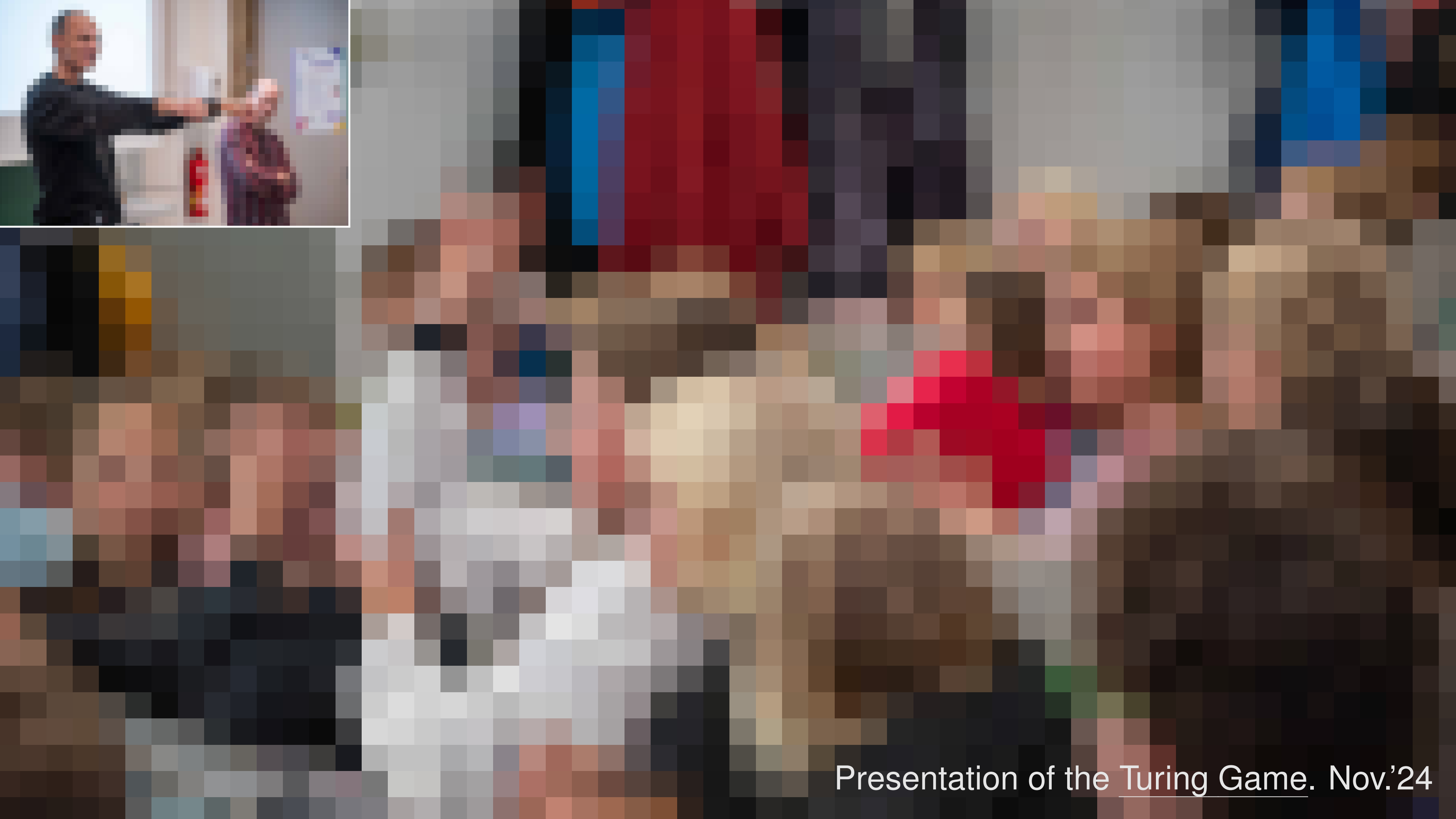
⑤ Watermarking Large Language Models


A
B
C
?

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.

- ❑ The “Turing Test” was called “Imitation Game” in the original paper.
- ❑ The Turing Test does not explain how human intelligens “works”.
  (and was never intended to do)
- ❑ According to rumors, the proposal was not meant seriously.
- ❑ Turing risked his reputation with this proposal.



Presentation of the Turing Game. Nov.’24

The Turing Collective Test

The question is not whether machines think – but whether we trust them.
We want to define and implement the “Turing Collective Test” to evaluate
the democratic capacity of Artificial Intelligence and make it negotiable.

B. Stein, J. Kiesel, H. Schmidgen, M. Jakesch. April ’25

The Turing Collective Test

> The question is not whether machines think – but whether we trust them.
> We want to define and implement the “Turing Collective Test” to evaluate the democratic capacity of Artificial Intelligence and make it negotiable.
>
> B. Stein, J. Kiesel, H. Schmidgen, M. Jakesch. April ’25

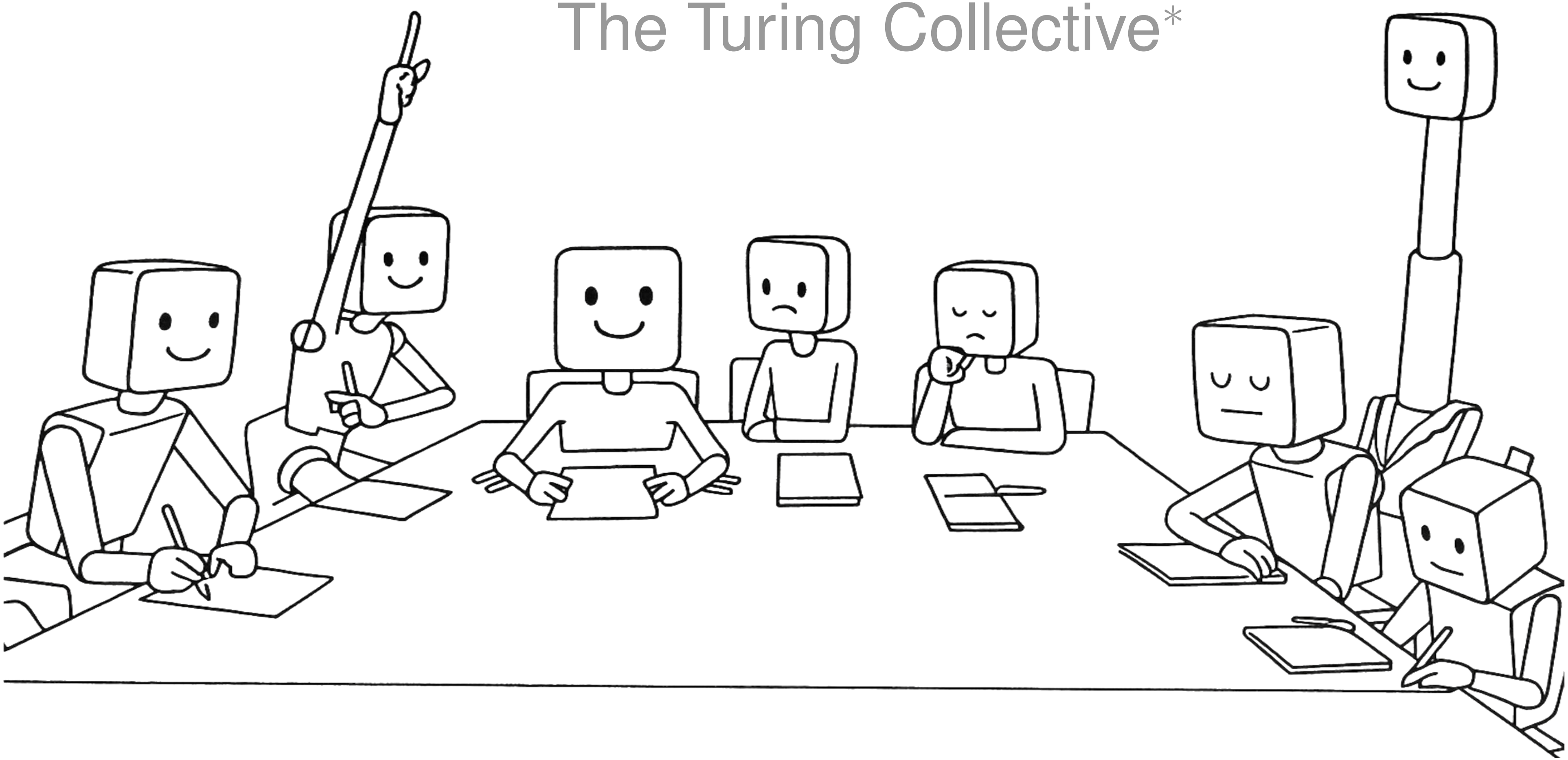
The components of the Turing collective test:

1. a hybrid collective C consisting of humans and AI agents,
2. a problem P, and
3. a human observer H.

C is given an amount of time to discuss P under the observation of H and propose a solution, which H either accepts or rejects. C passes the test if H accepts the solution proposed by C.

The Human Think Tank

The Turing Collective*



---

* Keyword “Denkkollektiv” (Fleck 1935): Entstehung und Entwicklung einer wissenschaftlichen Tatsache.

The Turing Collective Test

We envisage three stages for our test* :

1. Detection.
   Can people in a group discussion recognize AI agents posing as humans within the group and identify them as such?

2. Acceptance.
   When are people ready to accept AI agents into their communities and work with them?

3. Delegation.
   When are people willing to delegate decisions to collectives with AI agents?

---

* From the proposal “Der Turing-Kollektiv-Test”, Stein/Kiesel/Schmidgen/Jakesch. April ’25

Agenda

① Background on Large Language Models and Transformers

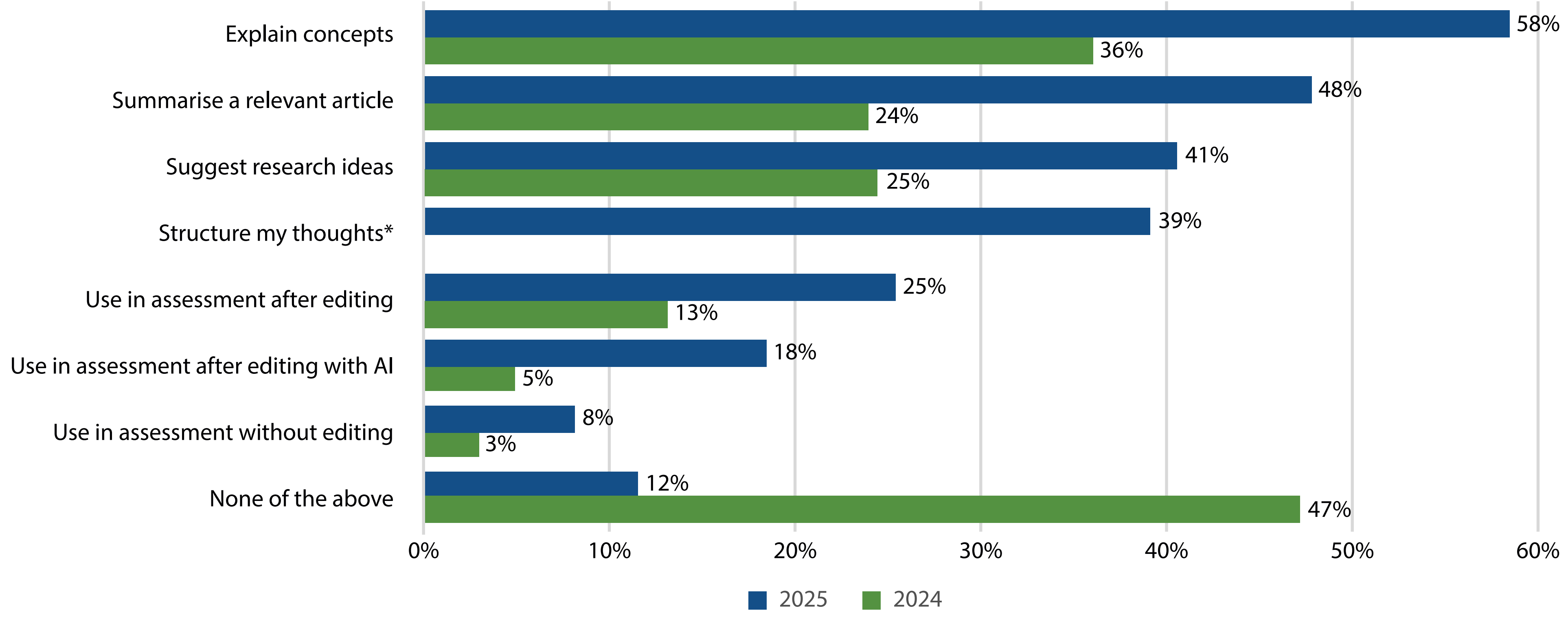
② Who is the Author? Generative LLM Authorship Verification

③ Turing X (interactive)

④ The Infobot Project – An LLM-based Teaching Prototype for Lectures

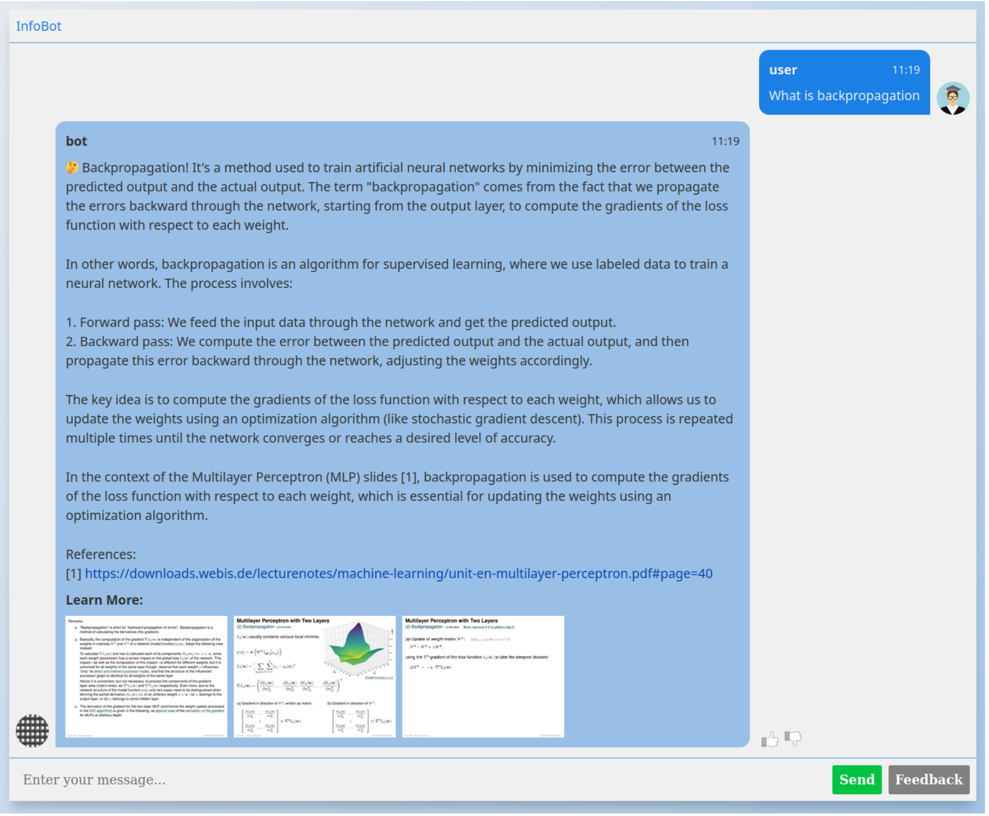
⑤ Watermarking Large Language Models

How Do Students Use GenAI*



* HEPI – Student Generative AI Survey (2025)

The Infobot Project



user 11:19

What is backpropagation

bot 11:19

🤔 Backpropagation! It's a method used to train artificial neural networks by minimizing the error between the predicted output and the actual output. The term "backpropagation" comes from the fact that we propagate the errors backward through the network, starting from the output layer, to compute the gradients of the loss function with respect to each weight.

In other words, backpropagation is an algorithm for supervised learning, where we use labeled data to train a neural network. The process involves:

1. Forward pass: We feed the input data through the network and get the predicted output.
2. Backward pass: We compute the error between the predicted output and the actual output, and then propagate this error backward through the network, adjusting the weights accordingly.

The key idea is to compute the gradients of the loss function with respect to each weight, which allows us to update the weights using an optimization algorithm (like stochastic gradient descent). This process is repeated multiple times until the network converges or reaches a desired level of accuracy.

In the context of the Multilayer Perceptron (MLP) slides [1], backpropagation is used to compute the gradients of the loss function with respect to each weight, which is essential for updating the weights using an optimization algorithm.

References:
[1] https://downloads.webis.de/lecturenotes/machine-learning/unit-en-multilayer-perceptron.pdf#page=40

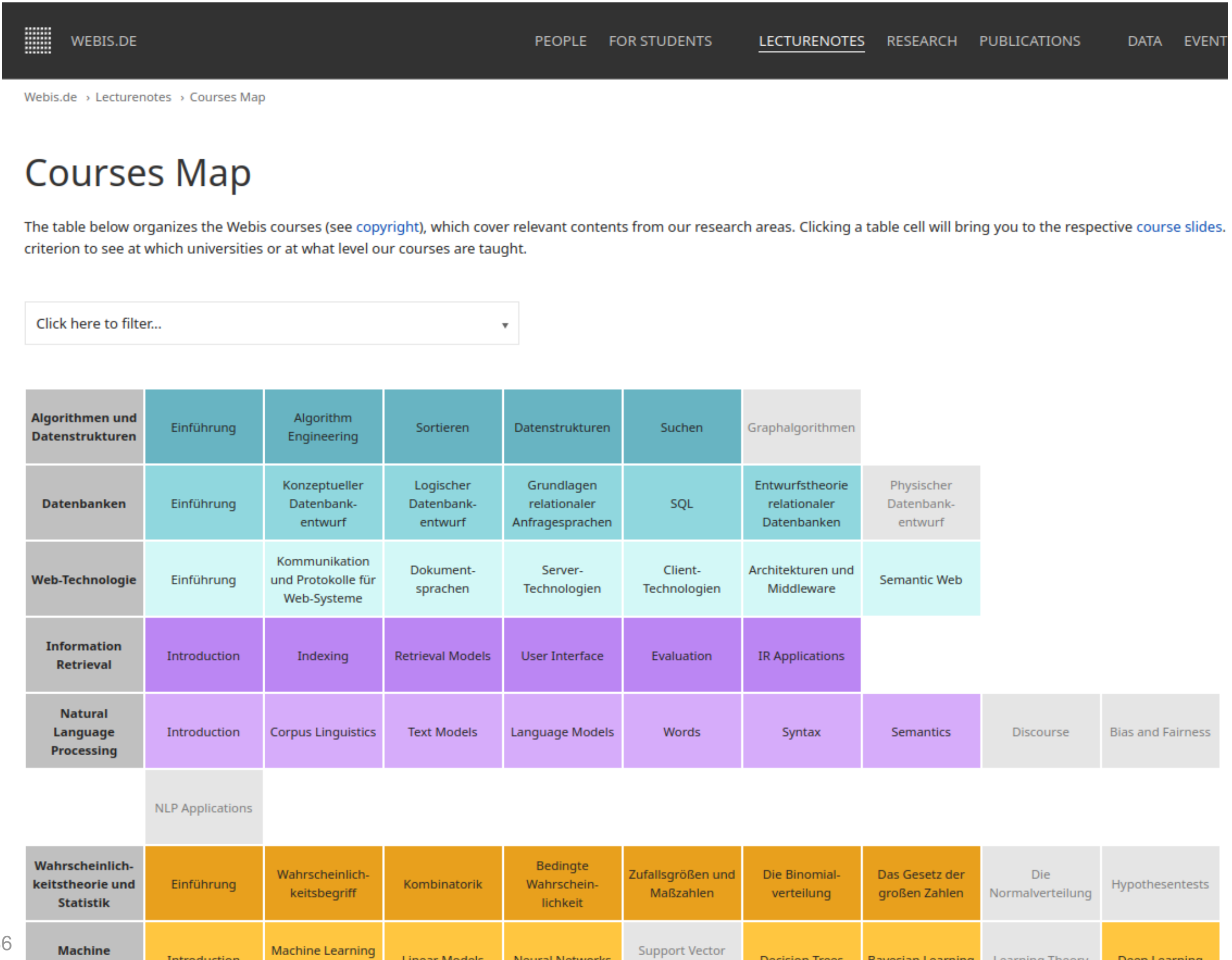
Learn More:

`infobot.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 (knowledge base construction)

lecturenotes.webis.de



Webis.de › Lecturenotes › Courses Map

Courses Map

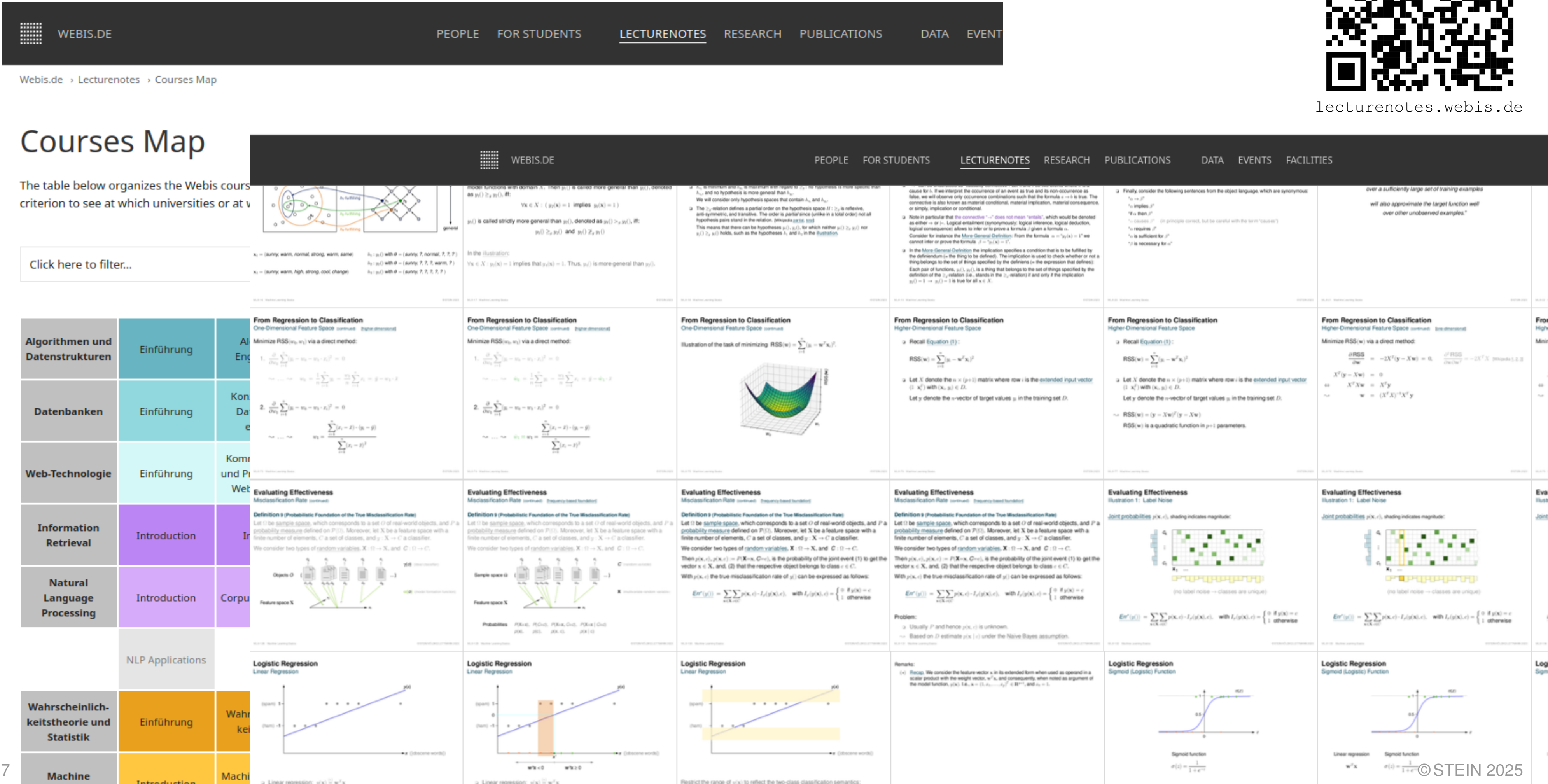
The table below organizes the Webis courses (see copyright), which cover relevant contents from our research areas. Clicking a table cell will bring you to the respective course slides. criterion to see at which universities or at what level our courses are taught.

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				

Course	Chapters	Units	Slides
Algorithms and Data Structures	5	17	926
Databases	6	15	756
Data Mining	5	12	381
Information Retrieval	6	18	1,020
Logics	5	18	663
Modeling KBS	6	21	741
Machine Learning	9	25	1,056
Natural Language Processing	9	19	770
Probability Theory and Statistics	8	26	853
Search	7	18	1,003
Language Tools	3	4	33
Web Technology	6	23	1,019
Σ	75	216	10,121

The Infobot Project (knowledge base construction)

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 w

Click here to filter...

Algorithmen und Datenstrukturen	Einführung	Al Eng
Datenbanken	Einführung	Kon Da e
Web-Technologie	Einführung	Komm und P Web
Information Retrieval	Introduction	In
Natural Language Processing	Introduction	Corpu
	NLP Applications	
Wahrscheinlichkeitstheorie und Statistik	Einführung	Wahr kei
Machine	Introduction	Machi

WEBIS.DE PEOPLE FOR STUDENTS LECTURENOTES RESEARCH PUBLICATIONS DATA EVENTS FACILITIES

From Regression to Classification
One-Dimensional Feature Space

From Regression to Classification
One-Dimensional Feature Space

From Regression to Classification
One-Dimensional Feature Space

From Regression to Classification
Higher-Dimensional Feature Space

From Regression to Classification
Higher-Dimensional Feature Space

From Regression to Classification
Higher-Dimensional Feature Space

Evaluating Effectiveness
Misclassification Rate

Evaluating Effectiveness
Misclassification Rate

Evaluating Effectiveness
Misclassification Rate

Evaluating Effectiveness
Misclassification Rate

Evaluating Effectiveness
Illustration 1: Label Noise

Evaluating Effectiveness
Illustration 1: Label Noise

Logistic Regression
Linear Regression

Logistic Regression
Linear Regression

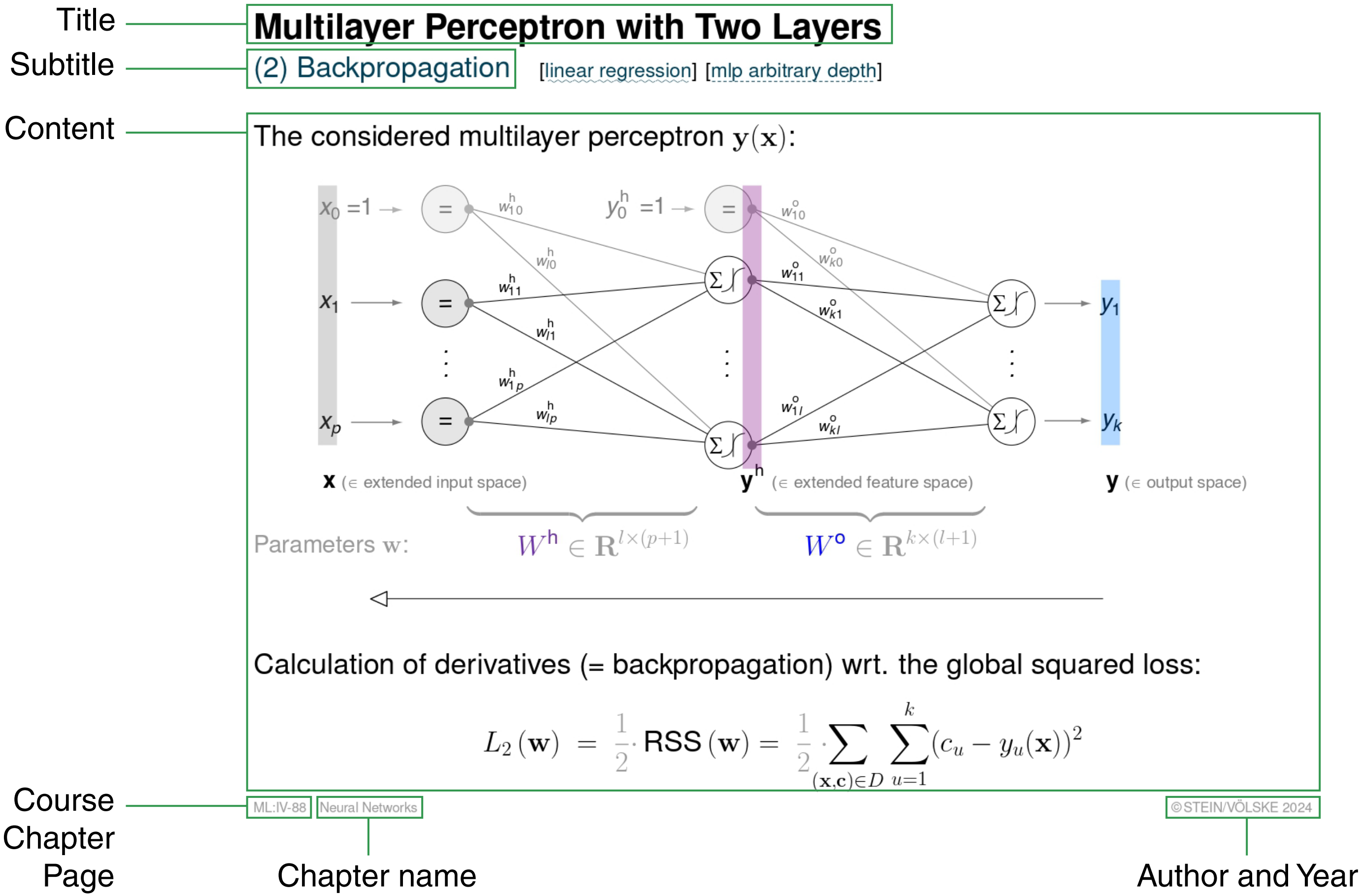
Logistic Regression
Linear Regression

Remarks:

Logistic Regression
Sigmoid (Logistic) Function

Logistic Regression
Sigmoid (Logistic) Function

The Infobot Project (knowledge base construction)



Title

Subtitle

Content

Course
Chapter
Page

Chapter name

Author and Year

Multilayer Perceptron with Two Layers

(2) Backpropagation

[linear regression] [mlp arbitrary depth]

The considered multilayer perceptron $\mathbf{y}(\mathbf{x})$:

Parameters $\mathbf{w}$: $W^h \in \mathbf{R}^{l\times(p+1)}$ $W^o \in \mathbf{R}^{k\times(l+1)}$

Calculation of derivatives (= backpropagation) wrt. the global squared loss:

$$L_2(\mathbf{w}) = \frac{1}{2}\cdot \text{RSS}(\mathbf{w}) = \frac{1}{2}\cdot \sum_{(\mathbf{x},\mathbf{c})\in D}\sum_{u=1}^{k}(c_u - y_u(\mathbf{x}))^2$$

ML:IV-88 Neural Networks ©STEIN/VÖLSKE 2024

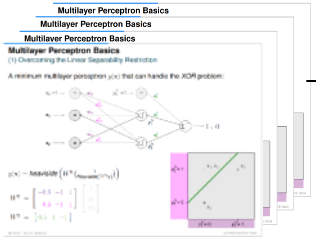
The Infobot Project (knowledge base construction)

```
\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



Multilayer Perceptron Basics

Multilayer Perceptron Basics

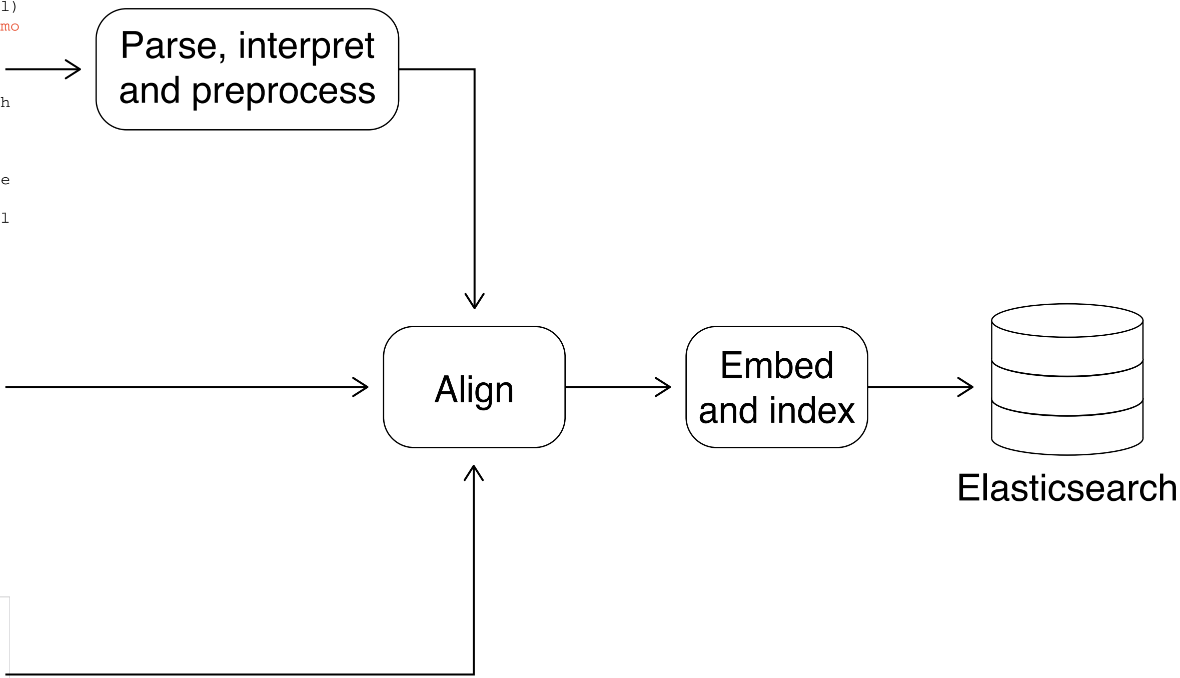
Multilayer Perceptron Basics

Multilayer Perceptron Basics

(1) Overcoming the Linear Separability Restriction

A minimum multilayer perceptron $y(x)$ that can handle the XOR problem:

Lecture slides



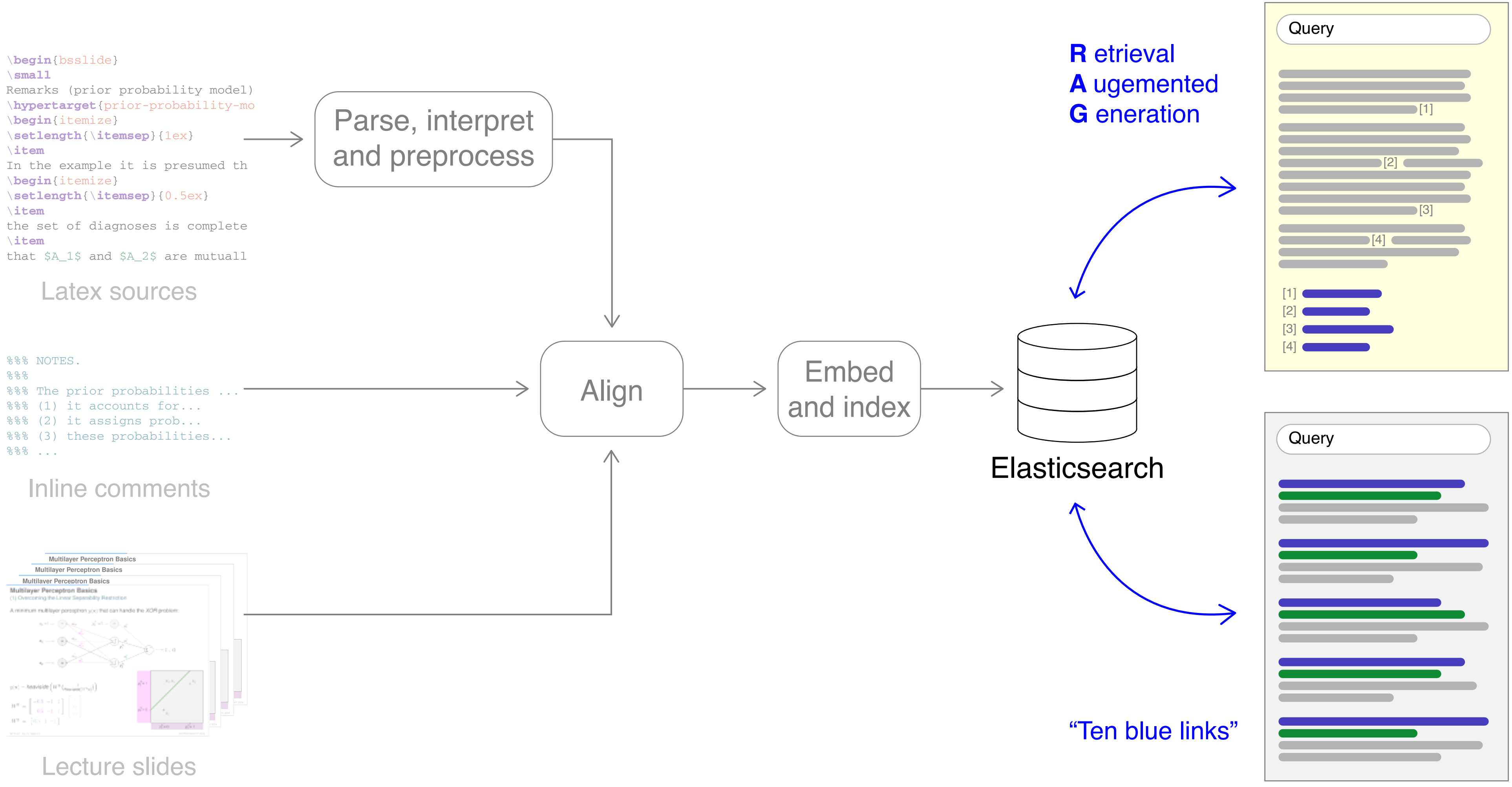
Parse, interpret and preprocess

Align

Embed and index

Elasticsearch

The Infobot Project (knowledge base construction)



```
\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

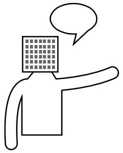
Embed and index

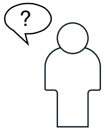
Elasticsearch

R etrieval
A ugemented
G eneration

“Ten blue links”

The Infobot Project (retrieval augmented generation)



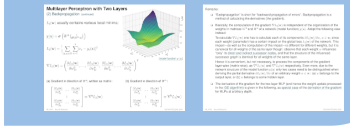


user 13:49

What is backpropagation?

bot 13:50

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



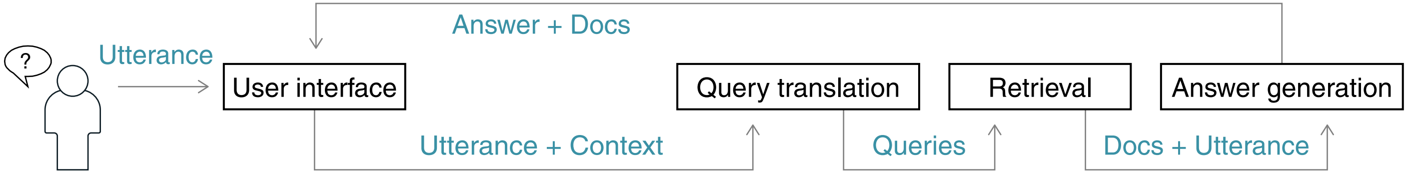
user 13:54

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 (retrieval augmented generation)



Utterance

User interface

Answer + Docs

Query translation

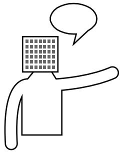
Retrieval

Answer generation

Utterance + Context

Queries

Docs + Utterance

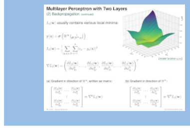


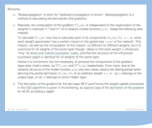
user 13:49

What is backpropagation?

bot 13:50

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





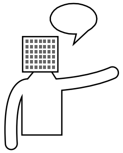
user 13:54

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 (retrieval augmented generation)

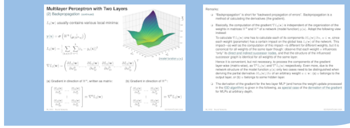


user 13:49

What is backpropagation?

bot 13:50

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

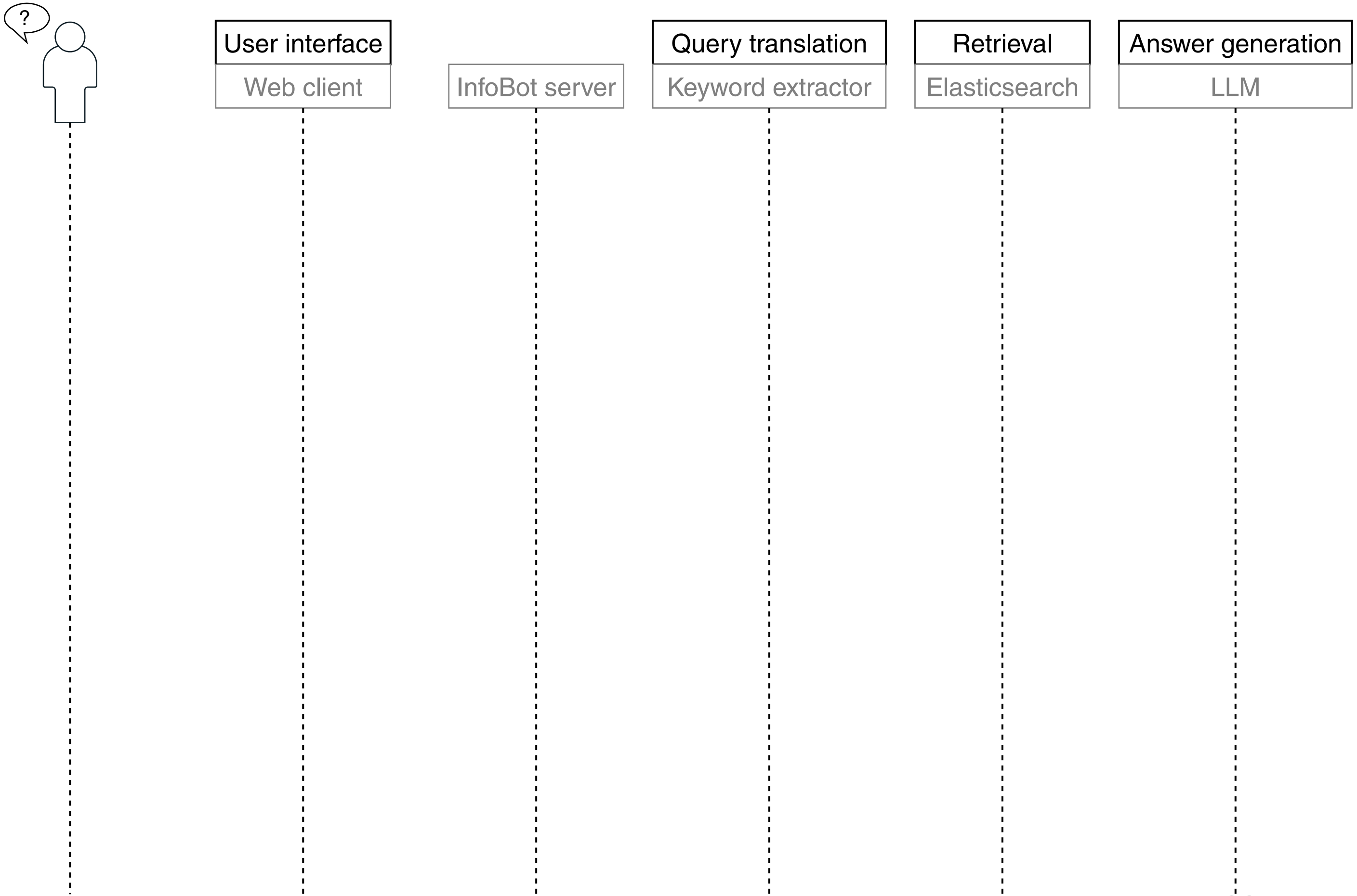


user 13:54

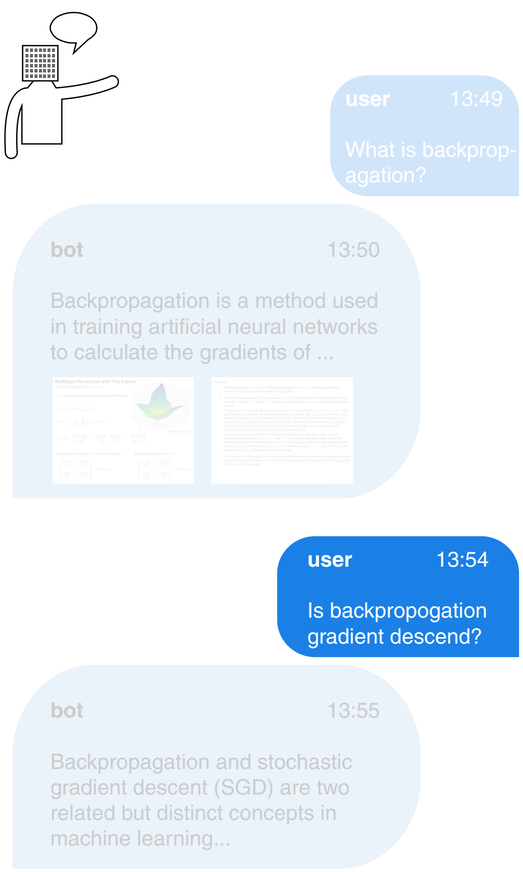
Is backpropogation gradient descend?

bot 13:55

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



The Infobot Project (retrieval augmented generation)



user 13:49

What is backpropagation?

bot 13:50

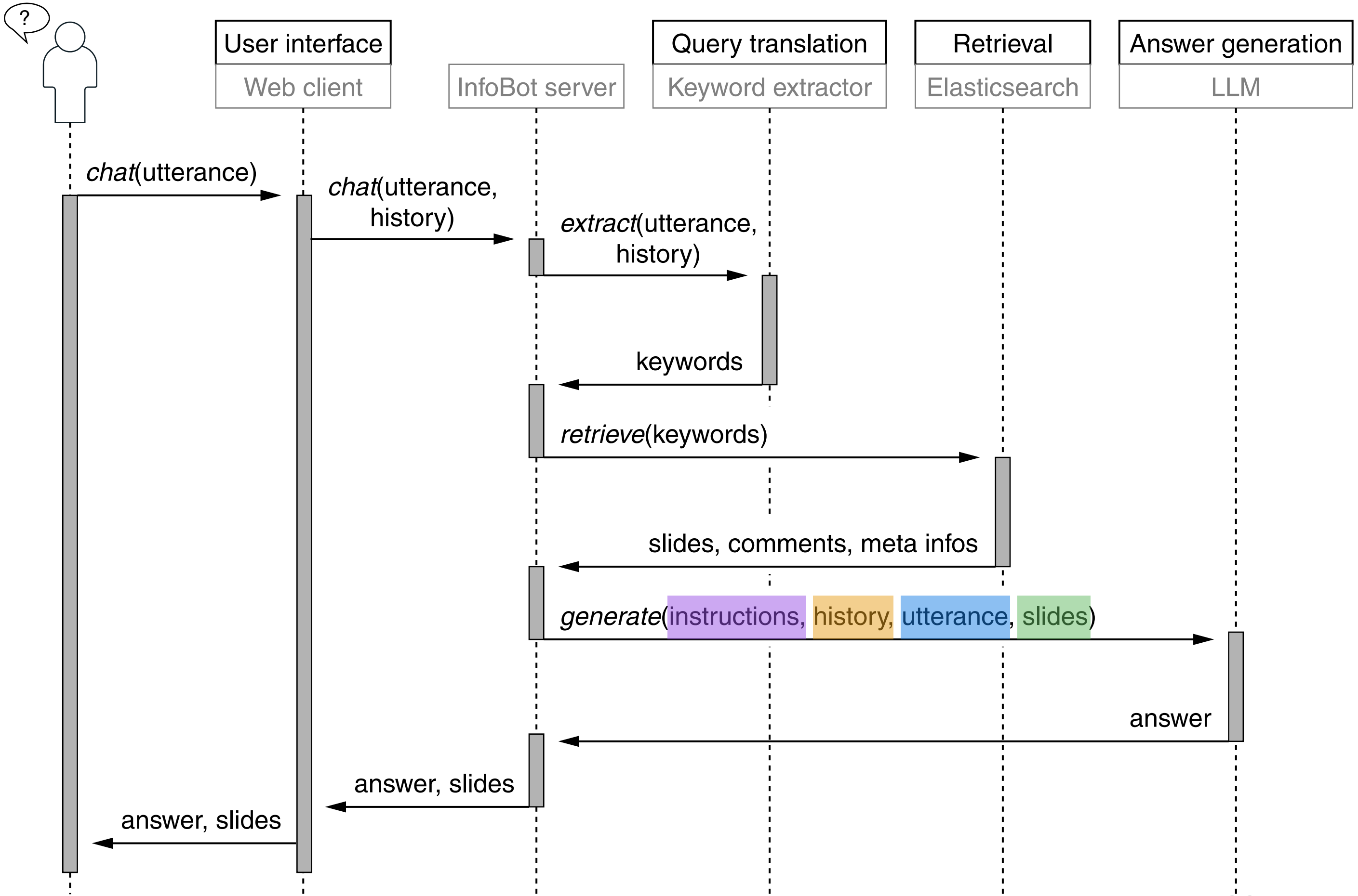
Backpropagation is a method used in training artificial neural networks to calculate the gradients of ...

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. Behavioral 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

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

The Infobot Project (background on retrieval, training, and evaluation)

- ❑ Query translation
  - (a) Keywords extracted with KeyBERT (`all-mpnet-base-v2`)
  - (b) Dense query vector with SBERT embeddings

- ❑ Retrieval model
  - (a) BM15 against slide title, subtitle and content
    Reranking: BM15 results weighted with keyword likelihood from KeyBERT
  - (b) k nearest neighbors

- ❑ Large language model
  - Meta Llama 3 (instruction-tuned)
  - 8 billion paramaters
  - 6-bit quantization

The Infobot Project (background on retrieval, training, and evaluation)

- Reinforcement learning with human feedback (RLHF)
  - Kahneman-Tversky optimization (KTO) based on manually created dataset with 100 questions

- Evaluation
  - Manually created dataset of 101 question-answer pairs and relevant slides
  - Cranfield-style IR experiments to analyze retrieval effectiveness
  - End-to-end evaluation with the Ragas framework:
    - Faithfulness: How factually consistent is the response with the retrieved slides?
    - Correctness: How factually consistent is the response with the ground-truth answer?
    - Relevancy: How relevant is the response for the user input?
  - Ablation studies and evaluation of different training and retrieval pipelines

The Infobot Project (background on retrieval, training, and evaluation)

- Reinforcement learning with human feedback (RLHF)
  - Kahneman-Tversky optimization (KTO) based on manually created dataset with 100 questions

- Evaluation
  - Manually created dataset of 101 question-answer pairs and relevant slides
  - Cranfield-style IR experiments to analyze retrieval effectiveness
  - End-to-end evaluation with the Ragas framework:
    - Faithfulness: How factually consistent is the response with the retrieved slides?
    - Correctness: How factually consistent is the response with the ground-truth answer?
    - Relevancy: How relevant is the response for the user input?
  - Ablation studies and evaluation of different training and retrieval pipelines

infobot.webis.de

Agenda

① Background on Large Language Models and Transformers

② Who is the Author? Generative LLM Authorship Verification

③ Turing X (interactive)

④ The Infobot Project – An LLM-based Teaching Prototype for Lectures

⑤ Watermarking Large Language Models


LLM
LLM

Watermarking Large Language Models

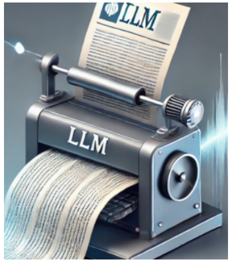
Distinguish between two scenarios:

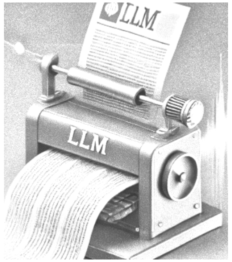
1. Generation-inherent Watermarking

   ⤳ Insert watermark during text generation

2. Post Watermarking

   ⤳ Insert watermark in existing text





[Demo]

Watermarking Large Language Models (generation-inherent)

Principle* :

1. Choose a secret, K, to generate unique seeds from token ids: $f_K(id) \rightarrow seed_{id}$
2. Randomly split vocabulary token-dependently, based on $seed_{id}$. ($\leadsto$ green list, red list)
3. **Generation:** When selecting a token at time $t{+}1$, prefer a list determined by token at t.
4. **Verification:** Analyze the list-dependent token occurrence probability, given a text.

---

* Original and variants: Kirchenbauer et al. (2023), Zhao et al. (2023), Lee et al. (2024), Lu et al. (2024)

Watermarking Large Language Models (generation-inherent)

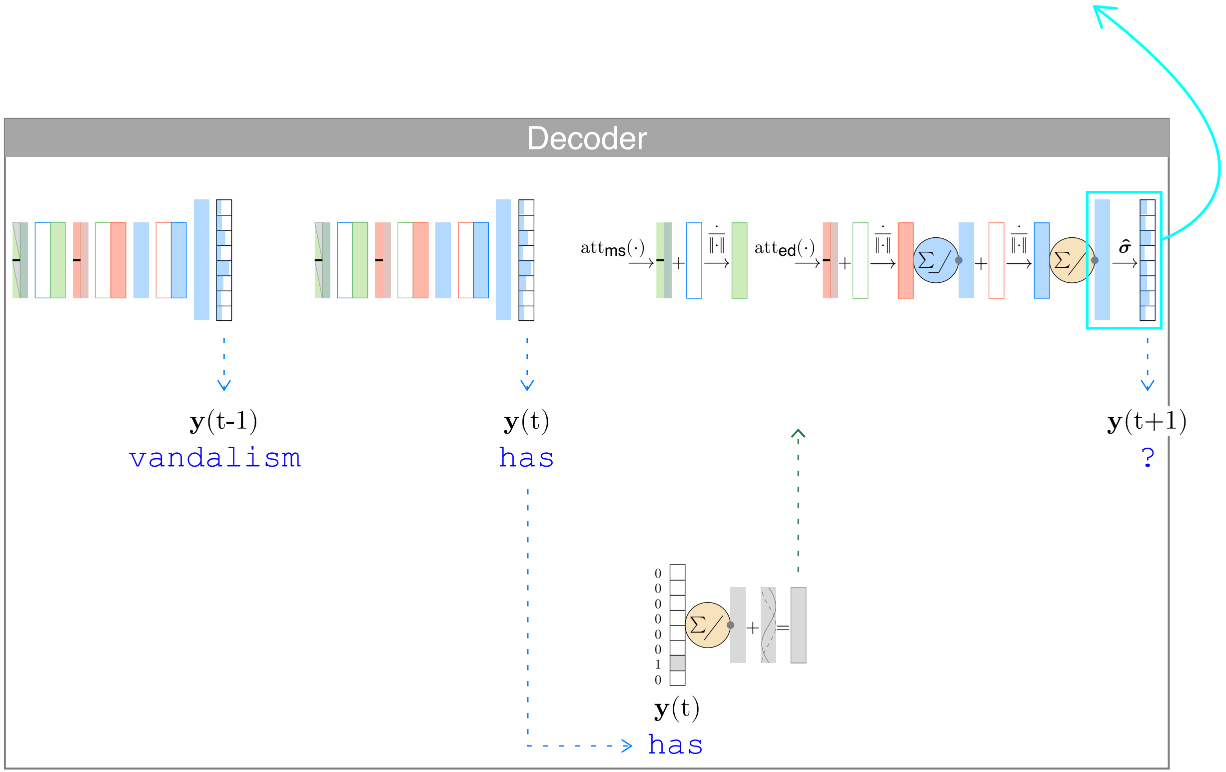
`Street art, once dismissed as mere vandalism, has evolved into a celebrated form of contemporary art.`

Watermarking Large Language Models (generation-inherent)

Street art, once dismissed as mere vandalism, has $\boxed{\mathbf{y}(t+1)}$ into a celebrated form of contemporary art

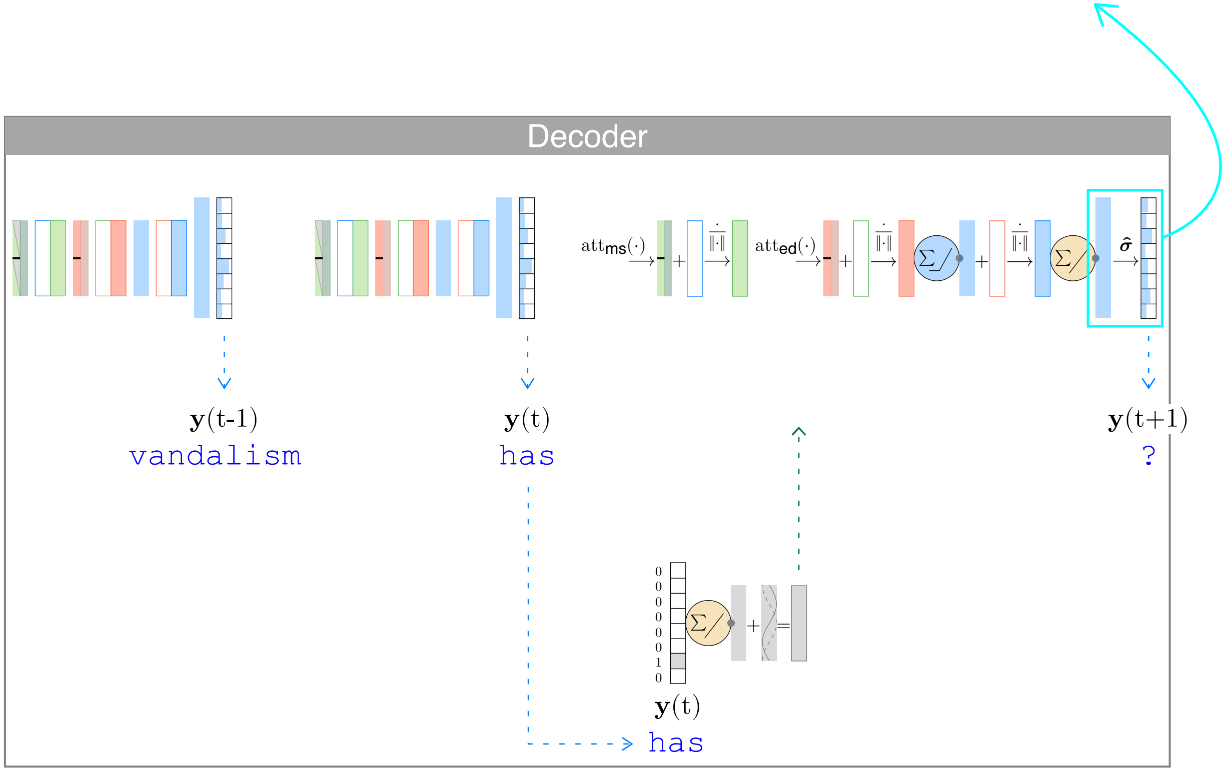
Watermarking Large Language Models (generation-inherent)

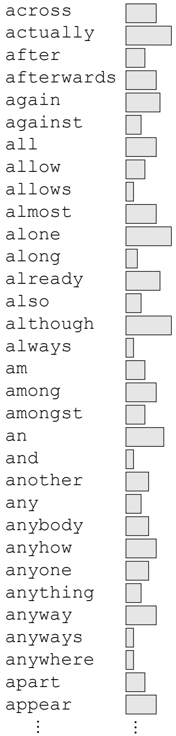
Street art, once dismissed as mere vandalism, has $\boxed{\mathbf{y}(\text{t+1})}$ into a celebrated form of contemporary art



Watermarking Large Language Models (generation-inherent)

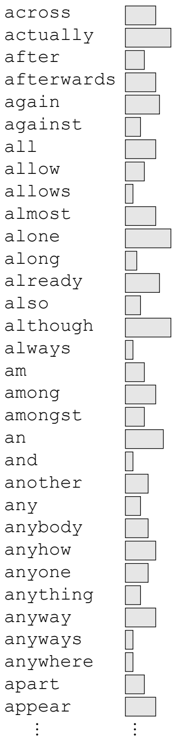
`Street art, once dismissed as mere vandalism, has` $\boxed{\mathbf{y}(\mathrm{t{+}1})}$ `into a celebrated form of contemporary art`





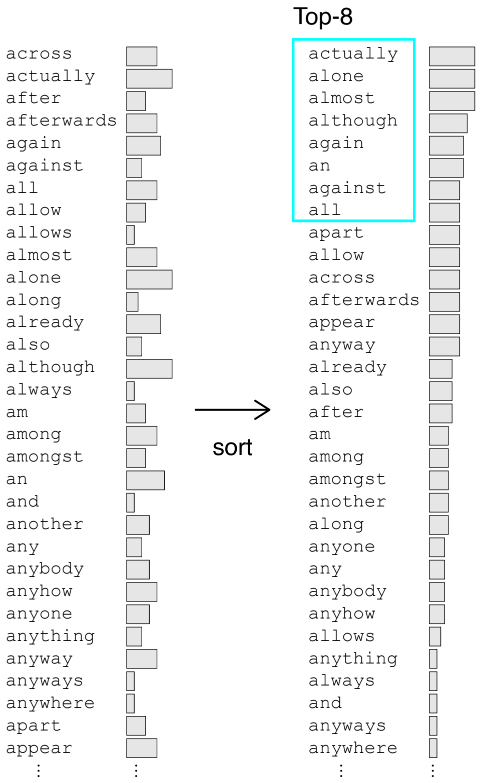
Watermarking Large Language Models (generation-inherent)

`Street art, once dismissed as mere vandalism, has` $\boxed{\mathbf{y}(t+1)}$ `into a celebrated form of contemporary art`



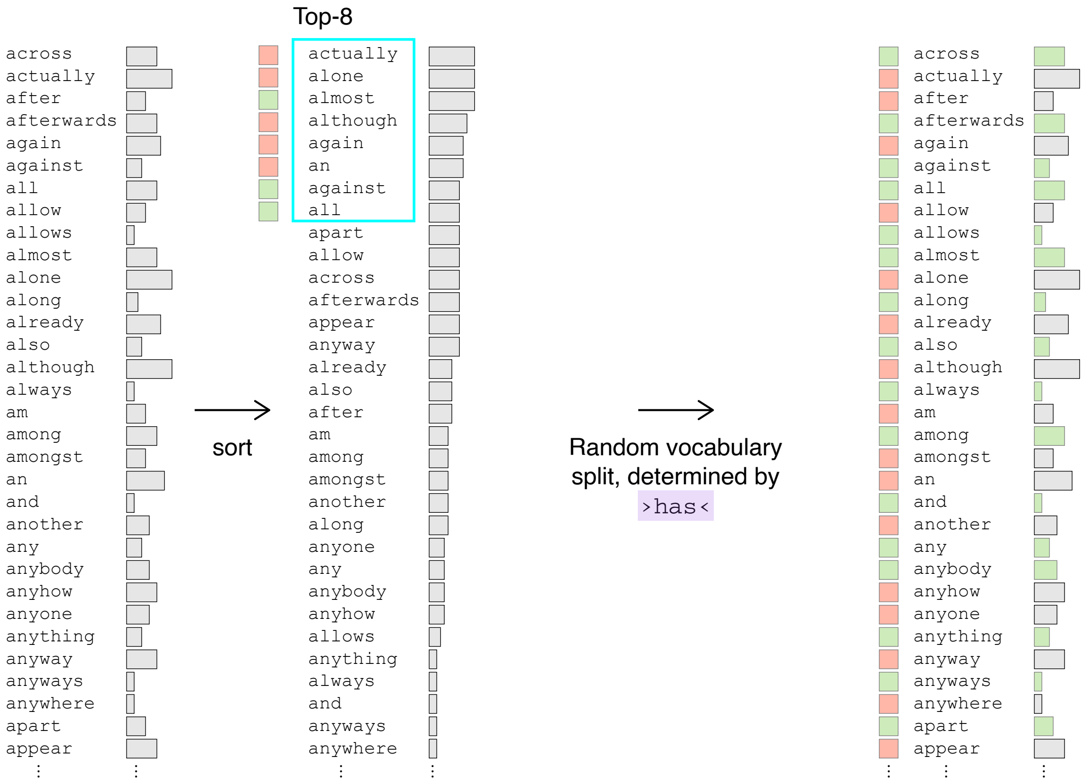
Watermarking Large Language Models (generation-inherent)

`Street art, once dismissed as mere vandalism, has` $\boxed{\mathbf{y}(t+1)}$ `into a celebrated form of contemporary art`



Watermarking Large Language Models (generation-inherent)

`Street art, once dismissed as mere vandalism, has` $\mathbf{y}(t+1)$ `into a celebrated form of contemporary art`



Top-8

across	actually		actually		🟩	across
across	🟥	actually		🟩	across	

Vocabulary
across
actually
after
afterwards
again
against
all
allow
allows
almost
alone
along
already
also
although
always
am
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anyway
anyways
anywhere
apart
appear
⋮

sort

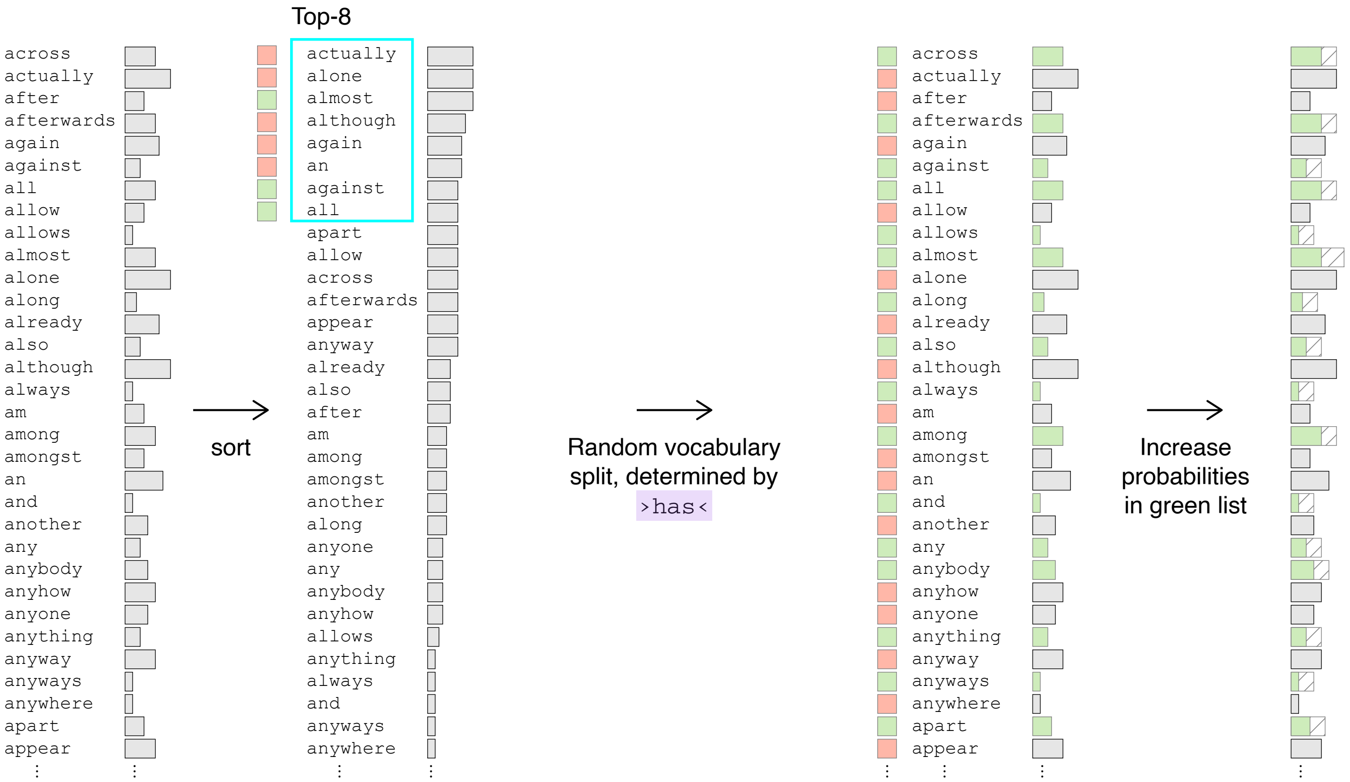
🟥	actually
🟥	alone
🟩	almost
🟥	although
🟥	again
🟥	an
🟩	against
🟩	all
	apart
	allow
	across
	afterwards
	appear
	anyway
	already
	also
	after
	am
	among
	amongst
	another
	along
	anyone
	any
	anybody
	anyhow
	allows
	anything
	always
	and
	anyways
	anywhere
	⋮

Random vocabulary split, determined by `>has<`

🟩	across
🟥	actually
🟥	after
🟩	afterwards
🟥	again
🟩	against
🟩	all
🟥	allow
🟩	allows
🟩	almost
🟥	alone
🟩	along
🟥	already
🟩	also
🟥	although
🟩	always
🟥	am
🟩	among
🟥	amongst
🟥	an
🟩	and
🟥	another
🟩	any
🟩	anybody
🟥	anyhow
🟥	anyone
🟩	anything
🟥	anyway
🟩	anyways
🟥	anywhere
🟩	apart
🟥	appear
⋮	⋮

Watermarking Large Language Models (generation-inherent)

`Street art, once dismissed as mere vandalism, has` $\mathbf{y}(t+1)$ `into a celebrated form of contemporary art`



Top-8

sort

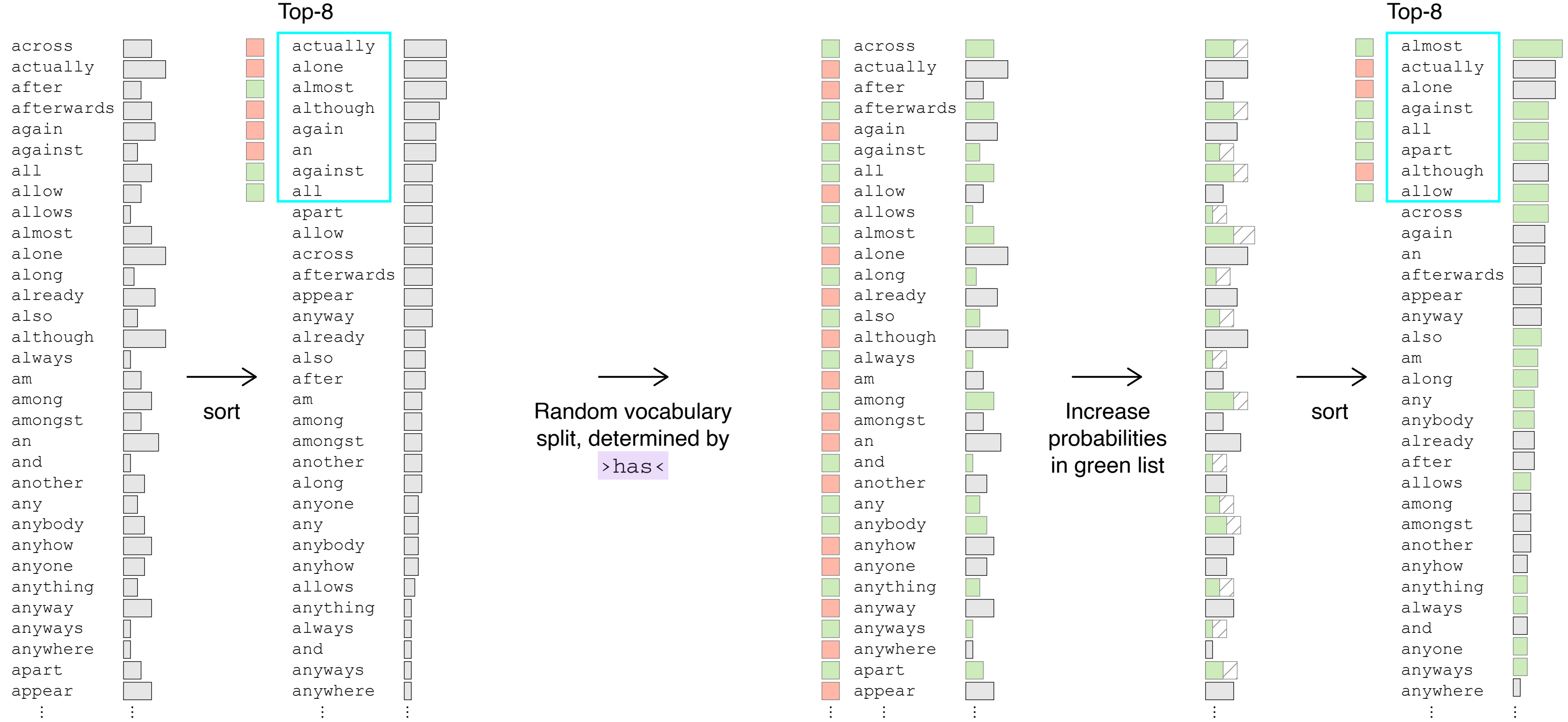
Random vocabulary split, determined by `>has<`

Increase probabilities in green list

Vocabulary	Top-8	Sorted	Split	Vocabulary
across	🟥	actually	🟩	across
actually	🟥	alone	🟥	actually
after	🟩	almost	🟥	after
afterwards	🟥	although	🟩	afterwards
again	🟥	again	🟥	again
against	🟥	an	🟩	against
all	🟩	against	🟩	all
allow	🟩	all	🟥	allow
allows		apart	🟩	allows
almost		allow	🟩	almost
alone		across	🟥	alone
along		afterwards	🟩	along
already		appear	🟥	already
also		anyway	🟩	also
although		already	🟥	although
always		also	🟩	always
am		after	🟥	am
among		am	🟩	among
amongst		among	🟥	amongst
an		amongst	🟥	an
and		another	🟩	and
another		along	🟥	another
any		anyone	🟩	any
anybody		any	🟩	anybody
anyhow		anybody	🟥	anyhow
anyone		anyhow	🟥	anyone
anything		allows	🟩	anything
anyway		anything	🟥	anyway
anyways		always	🟩	anyways
anywhere		and	🟥	anywhere
apart		anyways	🟩	apart
appear		anywhere	🟥	appear
⋮		⋮	⋮	⋮

Watermarking Large Language Models (generation-inherent)

`Street art, once dismissed as mere vandalism,` `has` $\mathbf{y}(t+1)$ `into a celebrated form of contemporary art`



sort

Random vocabulary split, determined by `>has<`

Increase probabilities in green list

sort

Watermarking Large Language Models (generation-inherent)

`Street art, once dismissed as mere vandalism, has evolved into a celebrated ...`

After watermarking:

`Street art, long regarded simply as vandalism, has transformed into a widely respected ...`

Watermarking Large Language Models (generation-inherent)

Street art, once dismissed as mere vandalism, has evolved into a celebrated ...

After watermarking:

Street art, long regarded simply as vandalism, has transformed into a widely respected ...

Watermarking Large Language Models (generation-inherent)

`Street art, once dismissed as mere vandalism, has evolved into a celebrated ...`

After watermarking:

`Street art, long regarded simply as vandalism, has transformed into a widely respected ...`

Possible setup:

- $\gamma = 0.5$ (green list size as fraction of token vocabulary)
- $\delta = 1.2$, 20% (factor or constant by which green list token probabilities are increased)

Without watermarking, the green tokens are binomially distributed, $B(n, p, k)$, with

- n = text sequence length T (time steps),
- $p = \gamma$,
- k = hits of tokens in green lists.

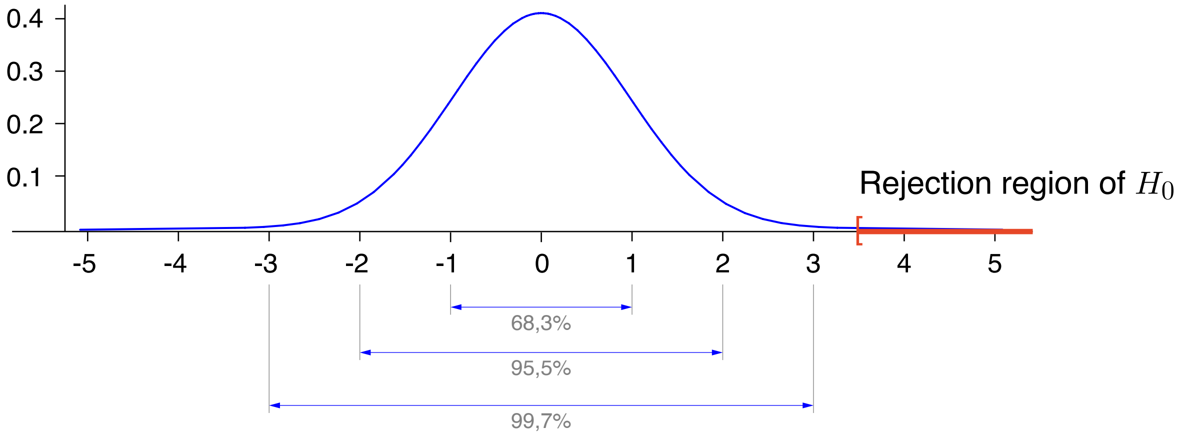
→ z-score: $z = \frac{k - \mu}{\sigma} = \frac{k - \gamma \cdot T}{\sqrt{T \cdot \gamma(1-\gamma)}}$

Watermarking Large Language Models (generation-inherent)

H_0: The text sequence has been generated with no token selection bias.

H_1: The text sequence has been generated with a green token preference.

z-scores:

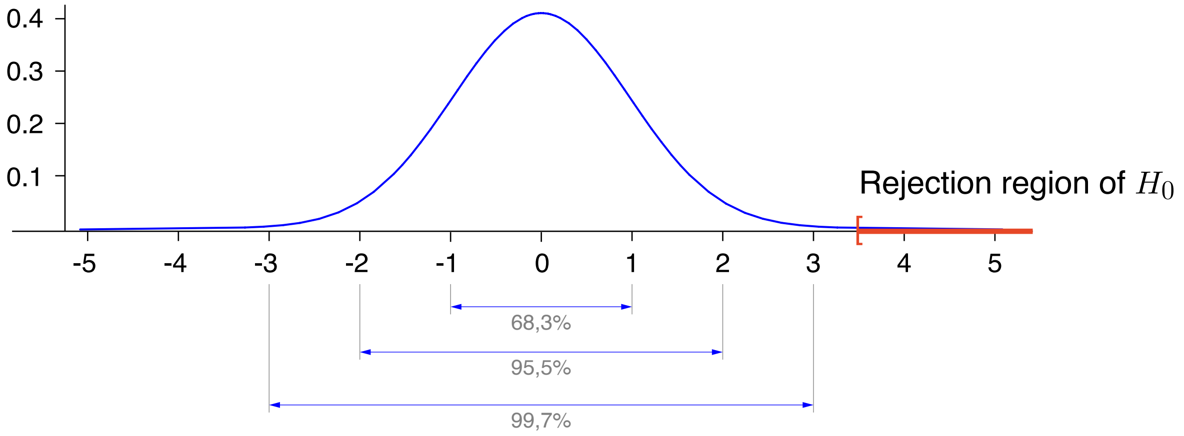


Watermarking Large Language Models (generation-inherent)

H_0: The text sequence has been generated with no token selection bias.

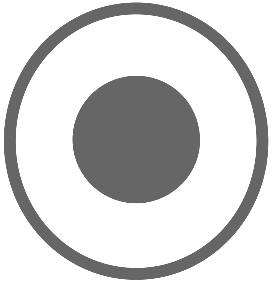
H_1: The text sequence has been generated with a green token preference.

z-scores:

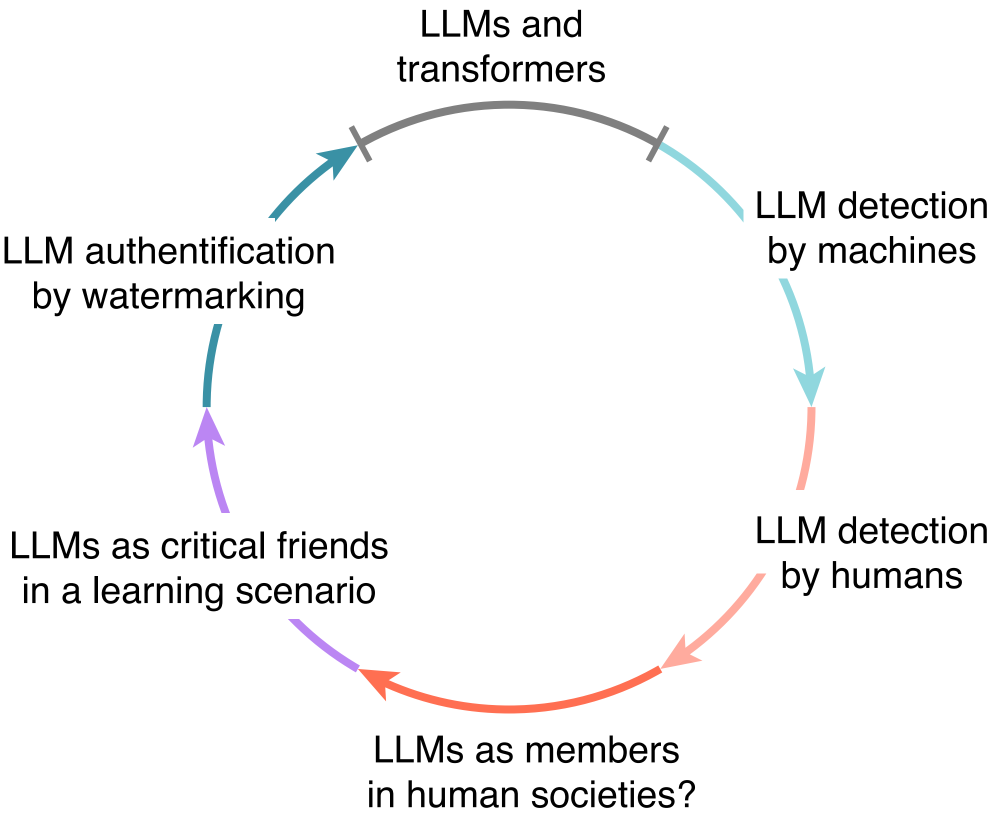


Example:

- ❑ Length T of text sequence = 200, $\gamma = 0.5$ (red and green lists have equal size)
- ❑ For $z = 3$ at least 121 green list tokens must be observed (instead of 100).



Recap of Our “Journey”



LLMs and transformers

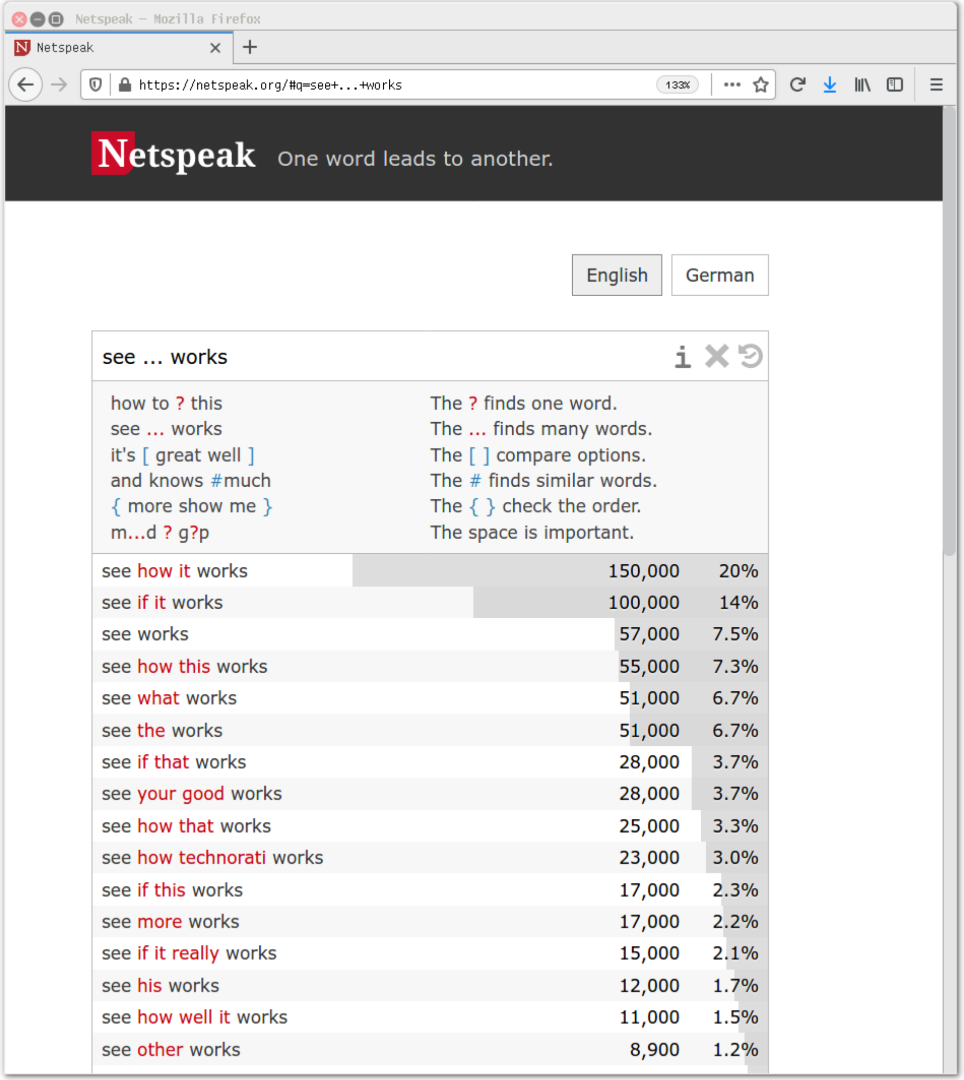
LLM detection by machines

LLM detection by humans

LLMs as members in human societies?

LLMs as critical friends in a learning scenario

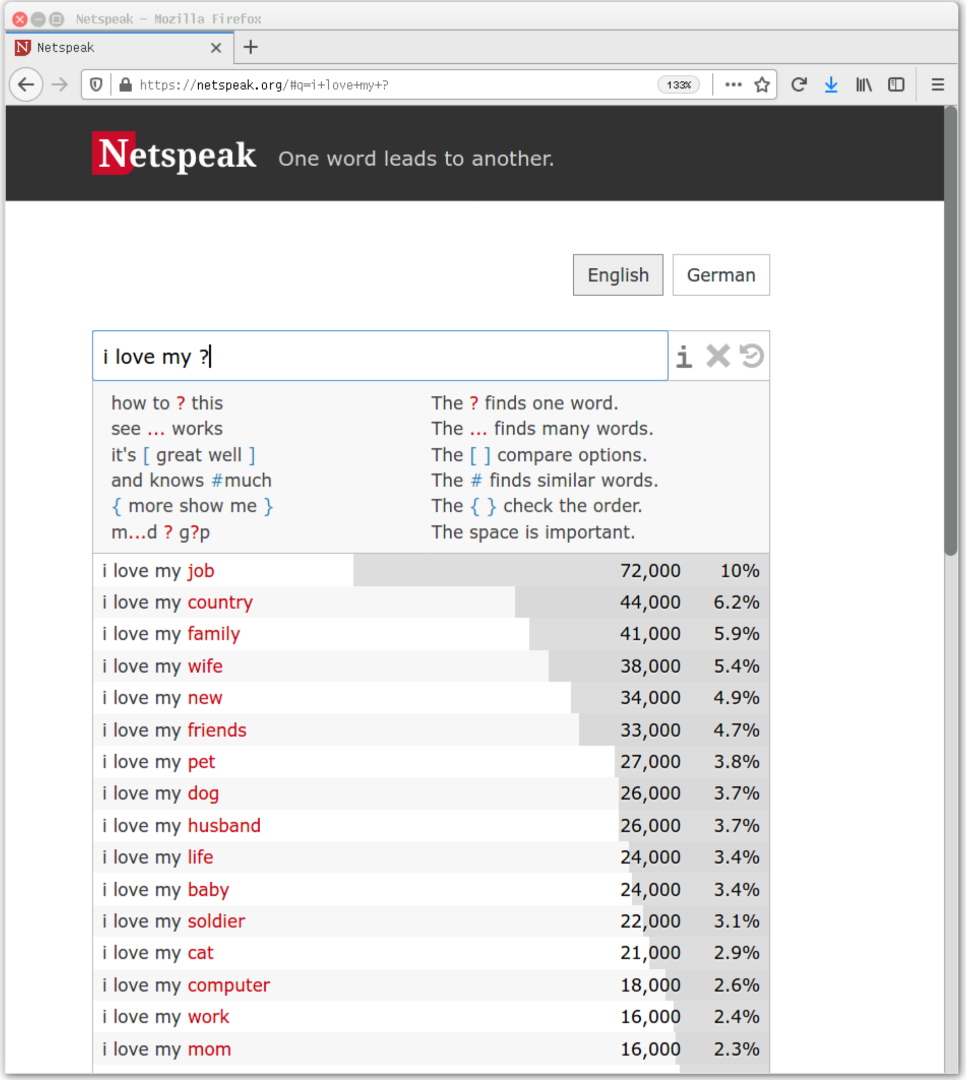
LLM authentification by watermarking



see ... works

how to ? this	The ? finds one word.
see ... works	The ... finds many words.
it's [great well]	The [ ] compare options.
and knows #much	The # finds similar words.
{ more show me }	The { } check the order.
m...d ? g?p	The space is important.

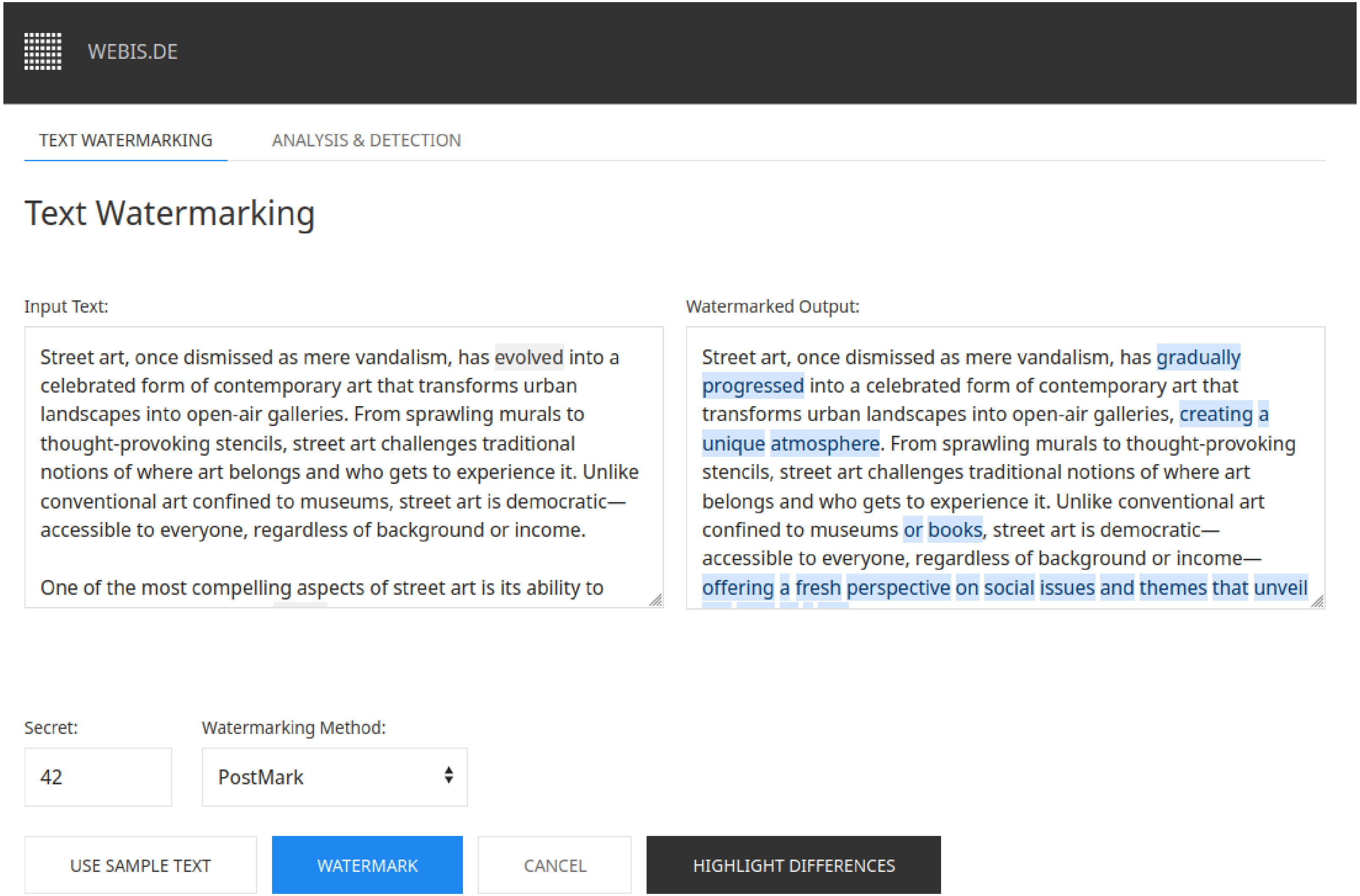
see how it works	150,000	20%
see if it works	100,000	14%
see works	57,000	7.5%
see how this works	55,000	7.3%
see what works	51,000	6.7%
see the works	51,000	6.7%
see if that works	28,000	3.7%
see your good works	28,000	3.7%
see how that works	25,000	3.3%
see how technorati works	23,000	3.0%
see if this 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%



i love my ?

how to ? this — The ? finds one word.
see ... works — The ... finds many words.
it's [great well] — The [ ] compare options.
and knows #much — The # finds similar words.
{ more show me } — The { } check the order.
m...d ? g?p — 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%



Text Watermarking

Input Text:

Street art, once dismissed as mere vandalism, has evolved into a celebrated form of contemporary art that transforms urban landscapes into open-air galleries. From sprawling murals to thought-provoking stencils, street art challenges traditional notions of where art belongs and who gets to experience it. Unlike conventional art confined to museums, street art is democratic—accessible to everyone, regardless of background or income.

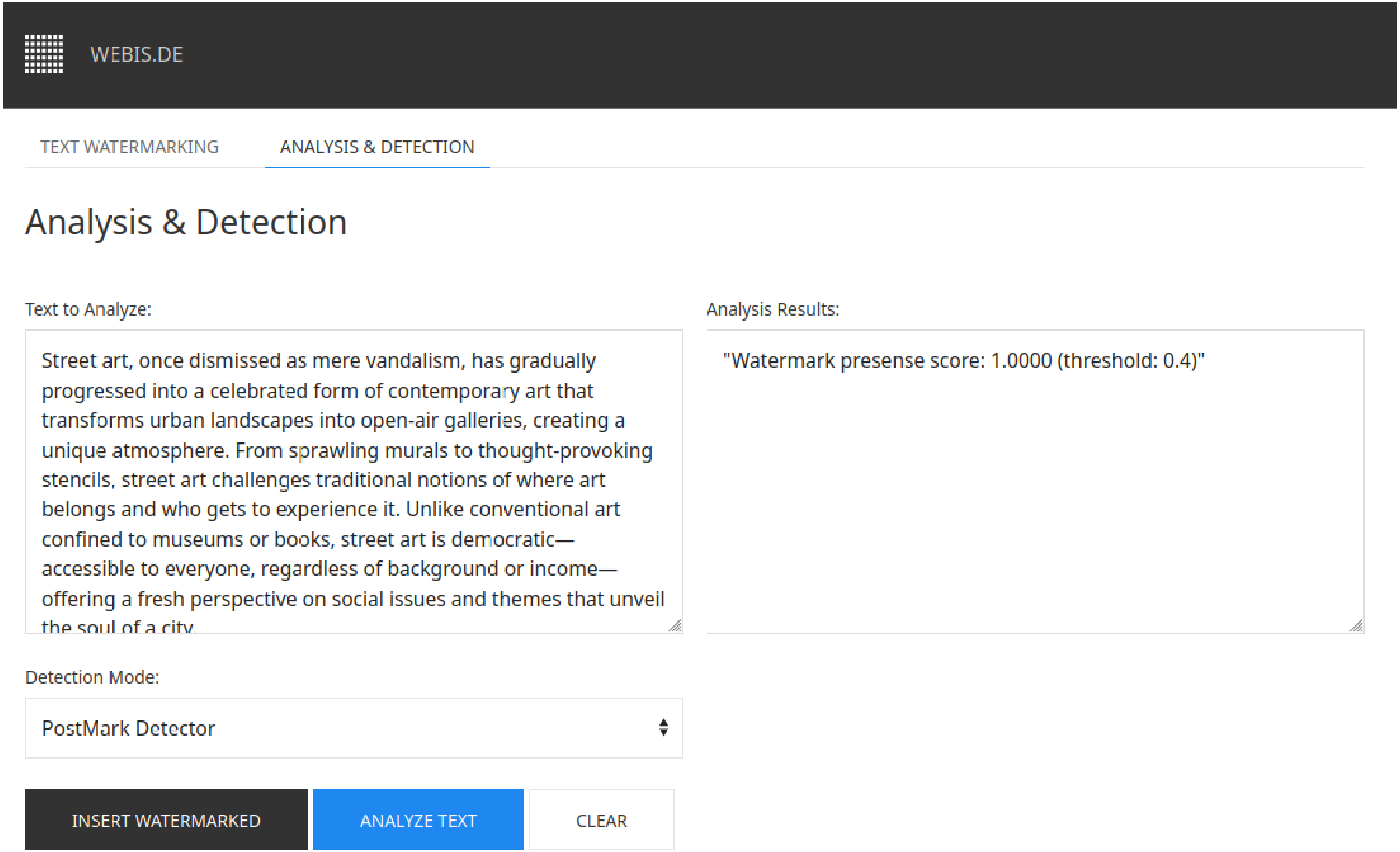
One of the most compelling aspects of street art is its ability to

Watermarked Output:

Street art, once dismissed as mere vandalism, has gradually progressed into a celebrated form of contemporary art that transforms urban landscapes into open-air galleries, creating a unique atmosphere. From sprawling murals to thought-provoking stencils, street art challenges traditional notions of where art belongs and who gets to experience it. Unlike conventional art confined to museums or books, street art is democratic—accessible to everyone, regardless of background or income—offering a fresh perspective on social issues and themes that unveil

Secret: 42

Watermarking Method: PostMark



Analysis & Detection

Text to Analyze:

Street art, once dismissed as mere vandalism, has gradually progressed into a celebrated form of contemporary art that transforms urban landscapes into open-air galleries, creating a unique atmosphere. From sprawling murals to thought-provoking stencils, street art challenges traditional notions of where art belongs and who gets to experience it. Unlike conventional art confined to museums or books, street art is democratic—accessible to everyone, regardless of background or income—offering a fresh perspective on social issues and themes that unveil the soul of a city

Analysis Results:

"Watermark presense score: 1.0000 (threshold: 0.4)"

Detection Mode: PostMark Detector