

# Complementing Course Materials with a Search-as-Learning Chatbot

A Prototype for Teaching Computer Science To University Students



Marcel Gohsen

Bauhaus-Universität  
Weimar

Research Group Intelligent Information Systems [webis.de]

- 1) Introduction
- 2) Concepts behind Infobot
- 3) Realization
- 4) Experiments and Results
- 5) Live Demo
- 6) Next Steps

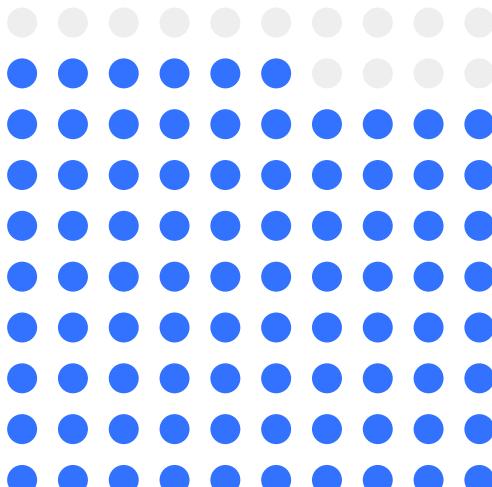
## 86% of students already use AI in their studies

### Percentage of students using AI in their studies

Question: How often do you use AI tools?

**86%**

of students claim to use  
AI in their studies

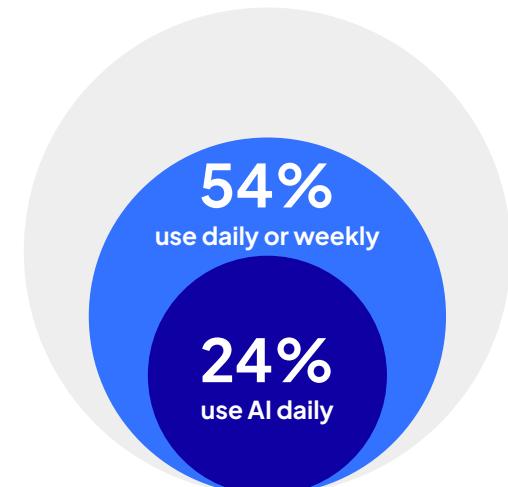


### Frequency of students using AI in their studies

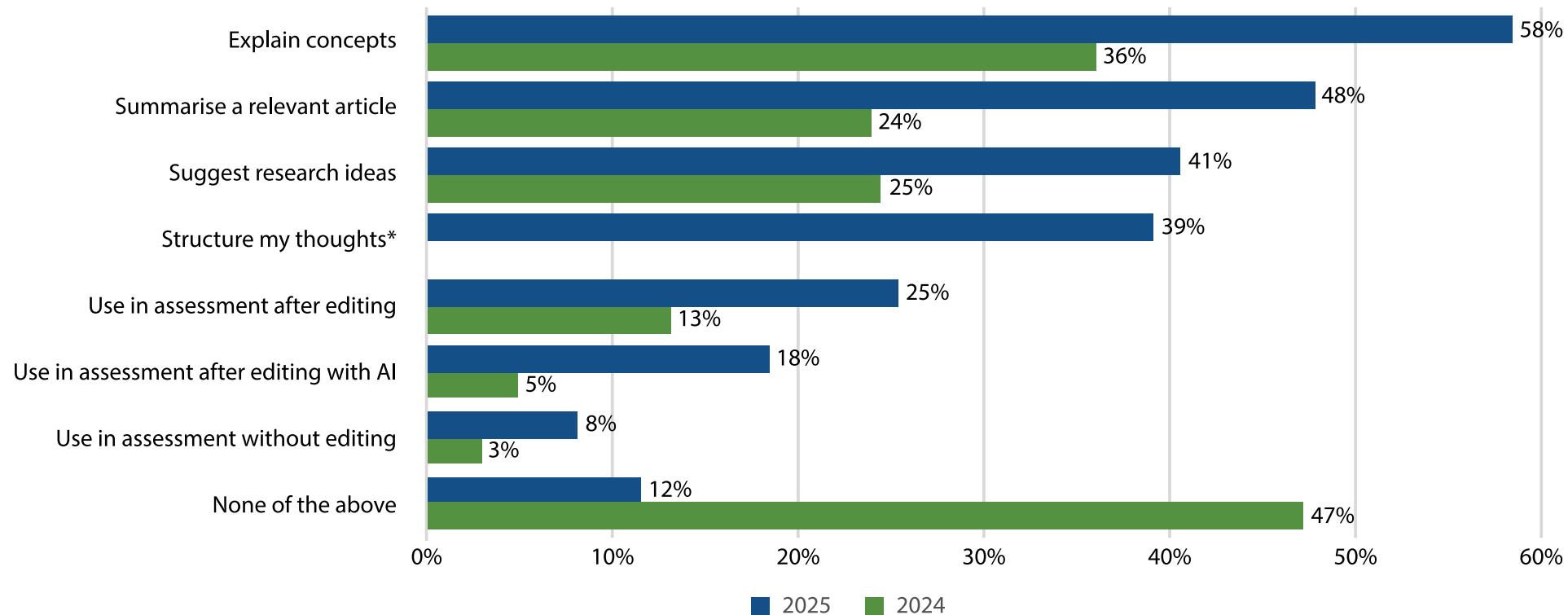
Question: How often do you use AI tools?

**54%**

of students use AI at least  
on a weekly basis



# How Do Students Use GenAI?



# “Learning” with ChatGPT



— How tall is the Eiffel Tower?

# “Learning” with ChatGPT



— How tall is the Eiffel Tower?

The Eiffel Tower is approximately 330 meters (1,083 feet) tall, including its antennas. Without the antennas, the structure itself is about 300 meters (984 feet) tall.



Factually correct!

# “Learning” with ChatGPT



— Why is CPU computation faster for deep learning?

# “Learning” with ChatGPT



— Why is CPU computation faster for deep learning?

CPU computation is generally faster than GPU computation for deep learning because CPUs are designed to handle a wide range of tasks and can switch between different tasks quickly and efficiently. This allows them to handle complex operations, such as data preprocessing and model training, more efficiently than GPUs, which are specifically designed for parallel processing of large amounts of data. Additionally, CPUs typically have larger cache memory and more advanced memory management techniques, which can improve the speed and efficiency of deep learning algorithms.



Contains hallucinations and lacks clarity!

# “Learning” with Infobot



— I have problems understanding backpropagation. Can you help me?

Don't worry. I'm here to help! Backpropagation is a powerful algorithm used in neural networks to train models by minimizing the error between predictions and actual outputs of a function. . . .

When you look at this slide from the Machine Learning course,  $W^h$  denotes the weight matrix for  $h$  layers . . .

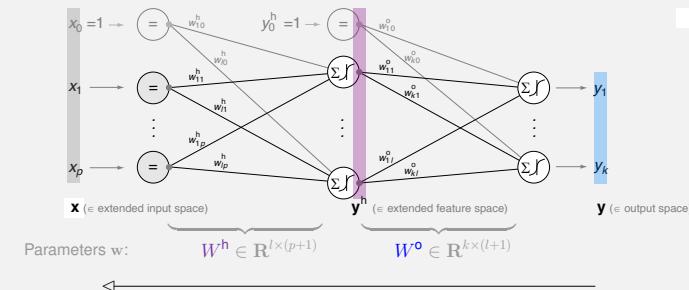
Please have a look at page 36 of the machine learning unit about multilayer perceptrons.



## Multilayer Perceptron with Two Layers

(2) Backpropagation [\[linear regression\]](#) [\[mlp arbitrary depth\]](#)

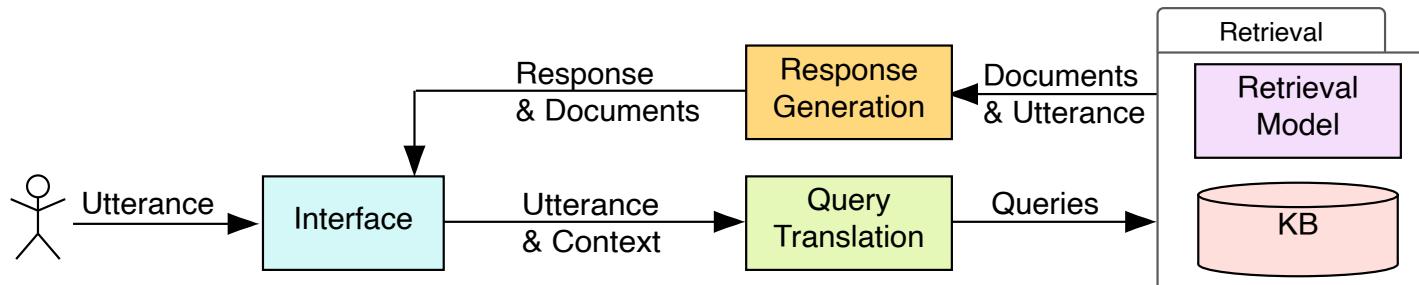
The considered multilayer perceptron  $\mathbf{y}(\mathbf{x})$ :



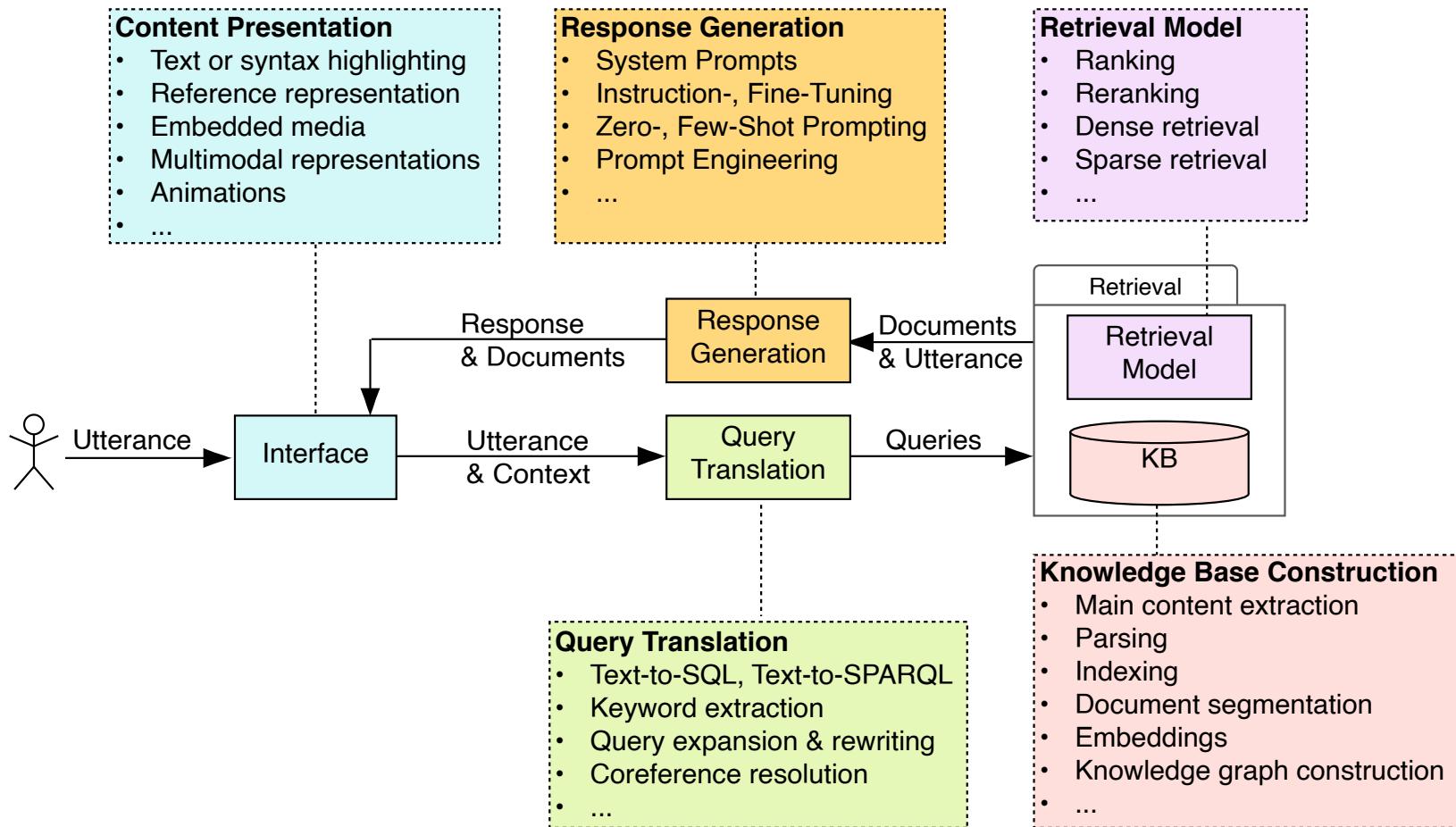
Calculation of derivatives (= backpropagation) wrt. the global squared loss:

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

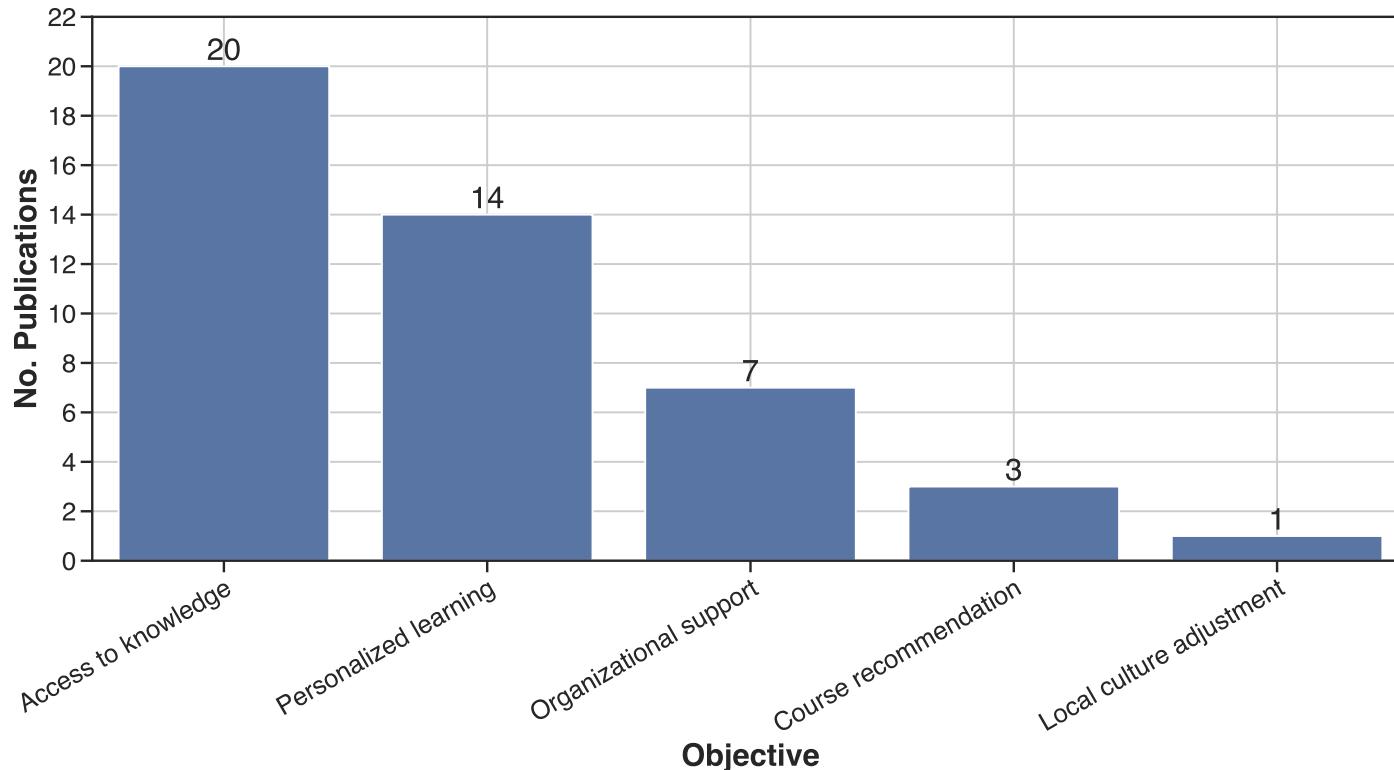
# Retrieval-Augmented Generation (RAG)



# Retrieval-Augmented Generation (RAG)



# RAG Systems in Education



Data from: Swacha and Gracel, 2025. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications.

# Requirements for Educational RAG

- ❑ Online course materials

Course materials should be online and accessible. (1/20)

- ❑ Links to sources

Responses should contain references to knowledge sources. (3/20)

- ❑ Multi-modal responses

Learning should be supported in multiple modalities. (2/20)

- ❑ Interactive content presentation

Responses should be presented in an intuitive, interactive and engaging way. (2/20)

- ❑ Openness

System should be available online to everyone. (1/20)

- ❑ Data privacy

Interaction data should be kept within university network. (3/20)

Currently, none of the systems fulfills all requirements.

## 2) Concepts behind Infobot

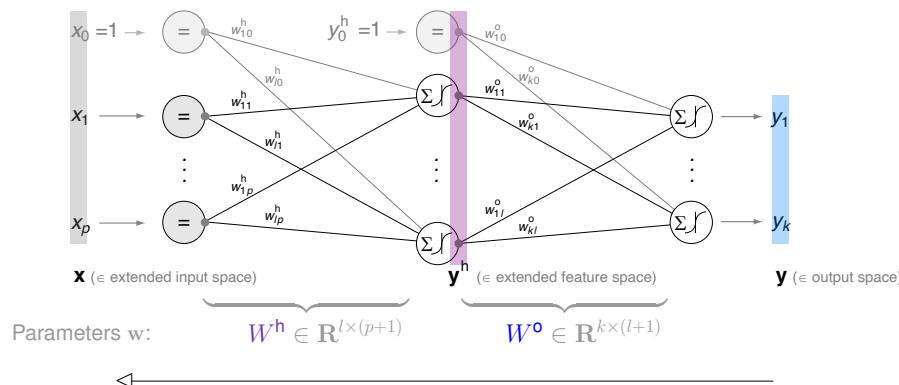
# Assumptions

## 1) Course Materials Contain the Answers

### Multilayer Perceptron with Two Layers

(2) Backpropagation [linear regression] [mlp\_arbitrary\_depth]

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



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

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

### Implications:

- Embed lecture slides in responses.
- Reference relevant course materials.
- Enable use as navigational retriever.
- Provide only grounded knowledge.

# Assumptions

## 2) Multimodal Presentation Supports Learning

Bimodally encoded stimuli increase recall from working memory. (Goolkasian and Foos, 2005)

- ❑ Increased recall: same object in written and spoken words.
- ❑ Decreased recall: different objects in written and spoken words.
- ❑ Worst recall: unimodal written words.

Multimodal encoding of semantically coherent stimuli enhances recall. (Delogu et al., 2009)

- ❑ Key: Semantically congruent and non-redundant stimuli.

Implications:

- ❑ Provide synchronized auditory and visual stimuli.
- ❑ Avoid written text-only presentations.

Goolkasian and Foos, 2005. Bimodal format effects in working memory.

Delogu et al., 2009. Semantic encoding in working memory: Is there a (multi)modality effect?

# Important Learning Theories

(from educational psychology)

## Behaviorism

- Learning as response to antecedent stimuli.
- Learning as response to consequences of stimuli (reward/punishment)

## Cognitivism

- Learning by copying someone's behaviour.
- Learning knowledge by head.

## Constructivism

- Learning by building on an individual's background knowledge.
- Learning by personal experience and unique interpretation.
- Learning by constructing personal knowledge themselves.

# Important Learning Theories

(from educational psychology)

## Behaviorism

- Learning as response to antecedent stimuli.
- Learning as response to consequences of stimuli (reward/punishment)

## Cognitivism

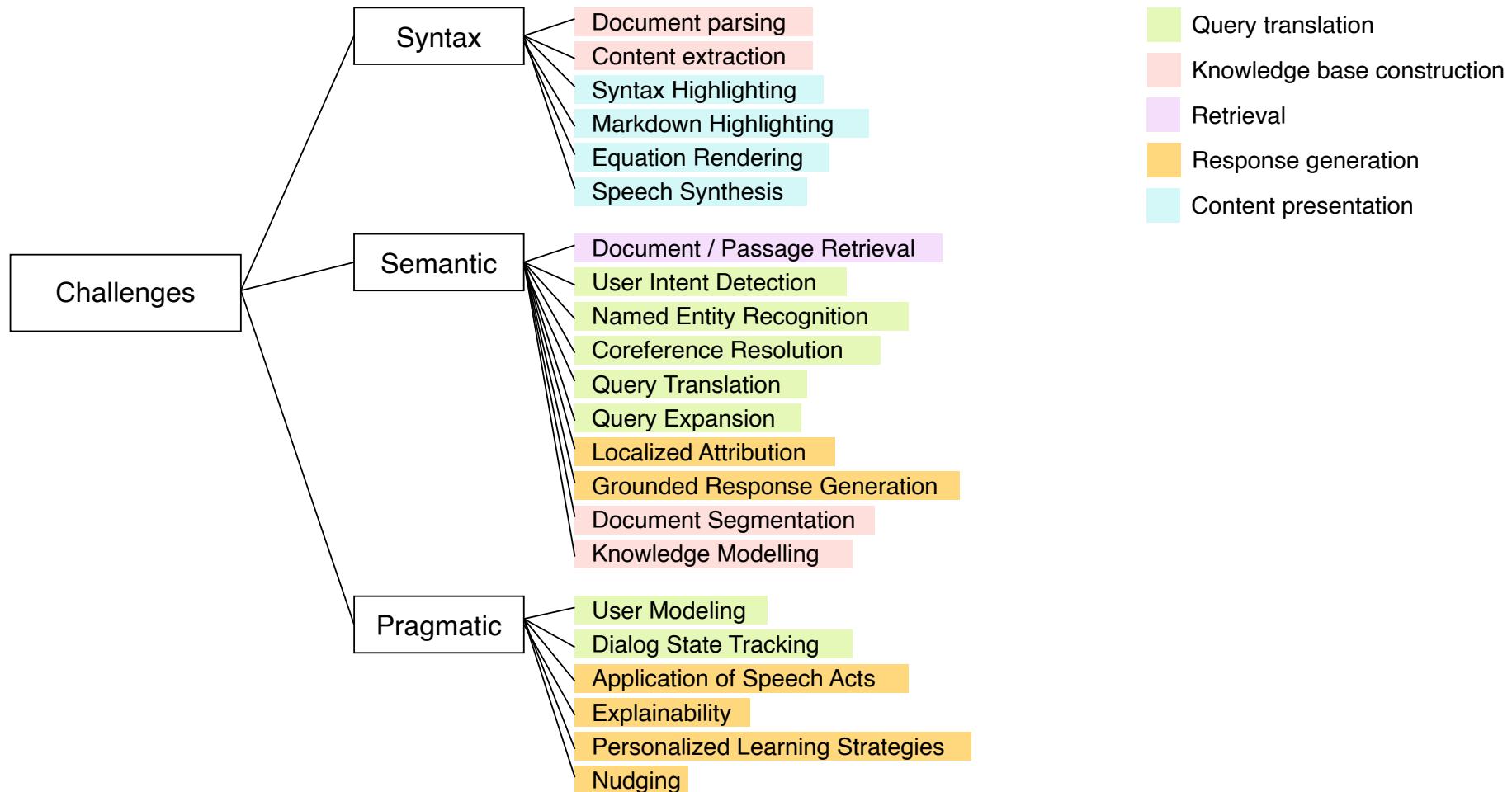
- Learning by copying someone's behaviour.
- Learning knowledge by head.

## Constructivism

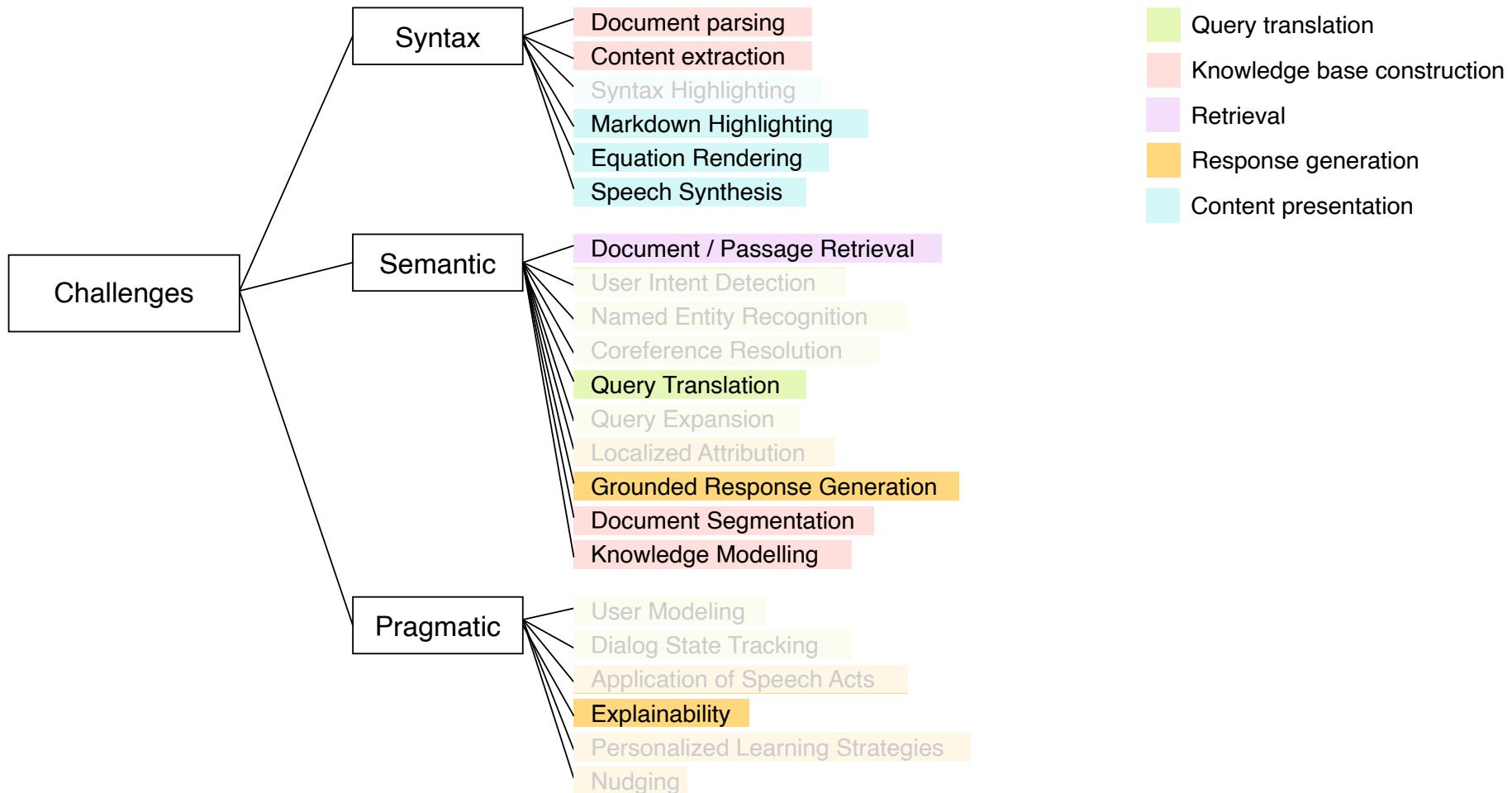
- Learning by building on an individual's background knowledge.
- Learning by personal experience and unique interpretation.
- Learning by constructing personal knowledge themselves.

Infobot: learning by uttering personal information needs through queries.

# Challenges in Developing a RAG System for Education



# Challenges in Developing a RAG System for Education



### 3) Realization of Infobot



## 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									
Wahrscheinlichkeitstheorie und Statistik	Einführung	Wahrscheinlichkeitsbegriff	Kombinatorik	Bedingte Wahrscheinlichkeit	Zufallsgrößen und Maßzahlen	Die Binomial-verteilung	Das Gesetz der großen Zahlen	Die Normalverteilung	Hypothesentests	
Machine	Machine Learning				Support Vector					



## Courses Map

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

Courses Map		
	WEBIS.DE	WEBIS.DE
<b>Algorithmen und Datenstrukturen</b>	<a href="#">Einführung</a>	<a href="#">Algo Eng</a>
<b>Datenbanken</b>	<a href="#">Einführung</a>	<a href="#">Kon Da e</a>
<b>Web-Technologie</b>	<a href="#">Einführung</a>	<a href="#">Kom und Pl</a>
<b>Information Retrieval</b>	<a href="#">Introduction</a>	<a href="#">Int</a>
<b>Natural Language Processing</b>	<a href="#">Introduction</a>	<a href="#">Corpu</a>
	<a href="#">NLP Applications</a>	
<b>Wahrscheinlichkeitstheorie und Statistik</b>	<a href="#">Einführung</a>	<a href="#">Wah</a>
<b>Machine</b>	<a href="#">Machine</a>	

**From Regression to Classification**

One-Dimensional Feature Space  $\mathbb{R}$  versus  $\mathbb{R}$ -dimensional

Minimize  $RSS(w_0, w_1)$  via a direct method:

$$1. \sum_{i=1}^n (w_0 + w_1 x_i - y_i)^2 = 0$$

$$w_0 + w_1 x_i - y_i = 0 \quad \forall i$$

$$w_0 + w_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0$$

$$2. \sum_{i=1}^n (w_0 + w_1 x_i - y_i - \bar{y})^2 = 0$$

$$w_0 + w_1 x_i - y_i - \bar{y} = 0 \quad \forall i$$

$$w_0 + w_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i - \bar{y} = 0$$

**From Regression to Classification**

One-Dimensional Feature Space  $\mathbb{R}$  versus  $\mathbb{R}$ -dimensional

Minimize  $RSS(w_0, w_1)$  via a direct method:

$$1. \sum_{i=1}^n (w_0 + w_1 x_i - y_i)^2 = 0$$

$$w_0 + w_1 x_i - y_i = 0 \quad \forall i$$

$$w_0 + w_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0$$

$$2. \sum_{i=1}^n (w_0 + w_1 x_i - y_i - \bar{y})^2 = 0$$

$$w_0 + w_1 x_i - y_i - \bar{y} = 0 \quad \forall i$$

$$w_0 + w_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i - \bar{y} = 0$$

**From Regression to Classification**

One-Dimensional Feature Space  $\mathbb{R}$  versus  $\mathbb{R}$ -dimensional

Minimize  $RSS(w_0, w_1)$  via a direct method:

Illustration of the task of minimizing  $RSS(w) = \sum_{i=1}^n (y_i - w^T x_i)^2$ .

**From Regression to Classification**

Higher-Dimensional Feature Space

Recall Equation (1):

$$RSS(w) = \sum_{i=1}^n (y_i - w^T x_i)^2$$

Let  $\mathbf{z}_i$  denote the  $n \times (p+1)$  matrix where row  $i$  is the extended input vector ( $\mathbf{x}_i$ ) with  $(\mathbf{x}_i, 1)$  as  $\mathbf{z}_i$ .

Let  $\mathbf{y}$  denote the  $n$ -vector of target values  $y_i$  in the training set  $D$ .

$RSS(w) = (\mathbf{y} - \mathbf{X}^T w)^T (\mathbf{y} - \mathbf{X}^T w)$

$RSS(w)$  is a quadratic function in  $p+1$  parameters.

**From Regression to Classification**

Higher-Dimensional Feature Space

Recall Equation (1):

$$RSS(w) = \sum_{i=1}^n (y_i - w^T x_i)^2$$

Let  $\mathbf{z}_i$  denote the  $n \times (p+1)$  matrix where row  $i$  is the extended input vector ( $\mathbf{x}_i$ ) with  $(\mathbf{x}_i, 1)$  as  $\mathbf{z}_i$ .

Let  $\mathbf{y}$  denote the  $n$ -vector of target values  $y_i$  in the training set  $D$ .

$RSS(w) = (\mathbf{y} - \mathbf{X}^T w)^T (\mathbf{y} - \mathbf{X}^T w)$

$RSS(w)$  is a quadratic function in  $p+1$  parameters.

**From Regression to Classification**

Higher-Dimensional Feature Space

Recall Equation (1):

$$RSS(w) = \sum_{i=1}^n (y_i - w^T x_i)^2$$

Let  $\mathbf{z}_i$  denote the  $n \times (p+1)$  matrix where row  $i$  is the extended input vector ( $\mathbf{x}_i$ ) with  $(\mathbf{x}_i, 1)$  as  $\mathbf{z}_i$ .

Let  $\mathbf{y}$  denote the  $n$ -vector of target values  $y_i$  in the training set  $D$ .

$RSS(w) = (\mathbf{y} - \mathbf{X}^T w)^T (\mathbf{y} - \mathbf{X}^T w)$

$RSS(w)$  is a quadratic function in  $p+1$  parameters.

**Evaluating Effectiveness**

Misclassification Rate  $\rightarrow$   $\text{Intrinsic-based measure}$

**Definition 1 (Probabilistic Foundation of the True Misclassification Rate)**

Let  $\mathcal{D}$  be a generic space, which corresponds to a set  $\mathcal{O}$  of real-world objects, and  $P$  a probability measure defined on  $\mathcal{P}(\mathcal{D})$ . Moreover, let  $\mathcal{X}$  be a feature space with a finite number of elements,  $\mathcal{C}$  a set of classes, and  $\mathcal{Y} \rightarrow \mathcal{C}$  a classifier.

We consider two types of random variables,  $X \rightarrow \mathcal{X}$ , and  $C \rightarrow \mathcal{C}$ .

These two types of random variables are the inputs of the classifier (1) to get the vector  $x \in \mathcal{X}$ , and (2) that the respective object belongs to class  $c \in \mathcal{C}$ .

With  $p(x, c)$  the true misclassification rate of  $x$ , it can be expressed as follows:

$$\text{Err}(x) = \sum_{c \in \mathcal{C}} p(x, c) - I_2(p(x, c)), \quad \text{with } I_2(p(x, c)) = \begin{cases} 0 & \text{if } p(x, c) < c \\ 1 & \text{otherwise} \end{cases}$$

We consider two types of random variables,  $X \rightarrow \mathcal{X}$ , and  $C \rightarrow \mathcal{C}$ .

These two types of random variables are the inputs of the classifier (1) to get the vector  $x \in \mathcal{X}$ , and (2) that the respective object belongs to class  $c \in \mathcal{C}$ .

With  $p(x, c)$  the true misclassification rate of  $x$ , it can be expressed as follows:

$$\text{Err}(x) = \sum_{c \in \mathcal{C}} p(x, c) - I_2(p(x, c)), \quad \text{with } I_2(p(x, c)) = \begin{cases} 0 & \text{if } p(x, c) < c \\ 1 & \text{otherwise} \end{cases}$$

Problem:

- Usually  $\mathcal{D}$  and hence  $p(x, c)$  is unknown.
- Based on  $\mathcal{D}$  estimate  $p(x, c)$  under the Naive Bayes assumption.

**Evaluating Effectiveness**

Illustration 1: Label Noise

Joint probabilities  $p(x, c)$ , shading indicates magnitude:

(no label noise  $\rightarrow$  classes are unique)

**Eval**

Joint probabilities  $p(x, c)$ , shading indicates magnitude:

(no label noise  $\rightarrow$  classes are unique)

**Eval**

Joint probabilities  $p(x, c)$ , shading indicates magnitude:

(no label noise  $\rightarrow$  classes are unique)

**Eval**

**Logistic Regression**

Linear Regression

$\text{Linear regression: } y(x) = w^T x$

**Logistic Regression**

Linear Regression

$\text{Linear regression: } y(x) = w^T x$

**Logistic Regression**

Linear Regression

$\text{Linear regression: } y(x) = w^T x$

**Logistic Regression**

Sigmoid (Logistic) Function

$\text{Sigmoid function: } \sigma(x) = \frac{1}{1 + e^{-x}}$

**Logistic Regression**

Sigmoid (Logistic) Function

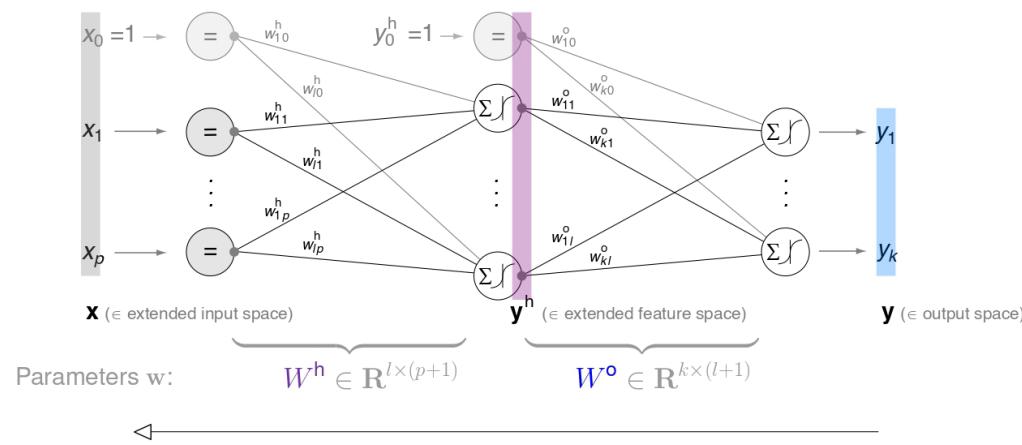
$\text{Sigmoid function: } \sigma(x) = \frac{1}{1 + e^{-x}}$

<b>Course</b>	<b>#Chapters</b>	<b>#Units</b>	<b>#Slides</b>
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
<b>Sum</b>	<b>75</b>	<b>216</b>	<b>10,121</b>

## Multilayer Perceptron with Two Layers

(2) Backpropagation [linear regression] [mlp arbitrary depth]

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



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

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

Title

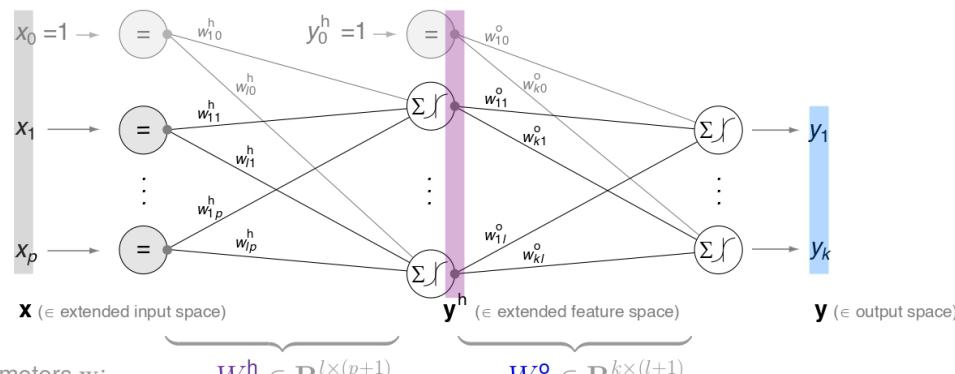
**Multilayer Perceptron with Two Layers**

Subtitle

## (2) Backpropagation

[linear regression] [mlp arbitrary depth]

Content

The considered multilayer perceptron  $y(x)$ :

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

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

Course  
Chapter  
Page

ML:IV-88 Neural Networks

Chapter Name

Lorem ipsum

©STEIN/VÖLSKE 2024

Author and Year

# Infobot (knowledge base construction)

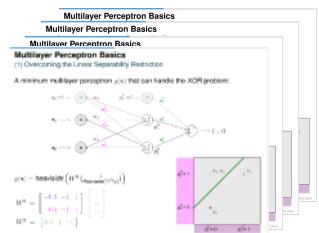
```
\begin{bsslide}
\small
Remarks (prior probability model)
\hypertarget{prior-probability-mo}
\begin{itemize}
\setlength{\itemsep}{1ex}
\item
In the example it is presumed th
\begin{itemize}
\setlength{\itemsep}{0.5ex}
\item
the set of diagnoses is complete
\item
that $A\_1$ and $A\_2$ are mutuall

```

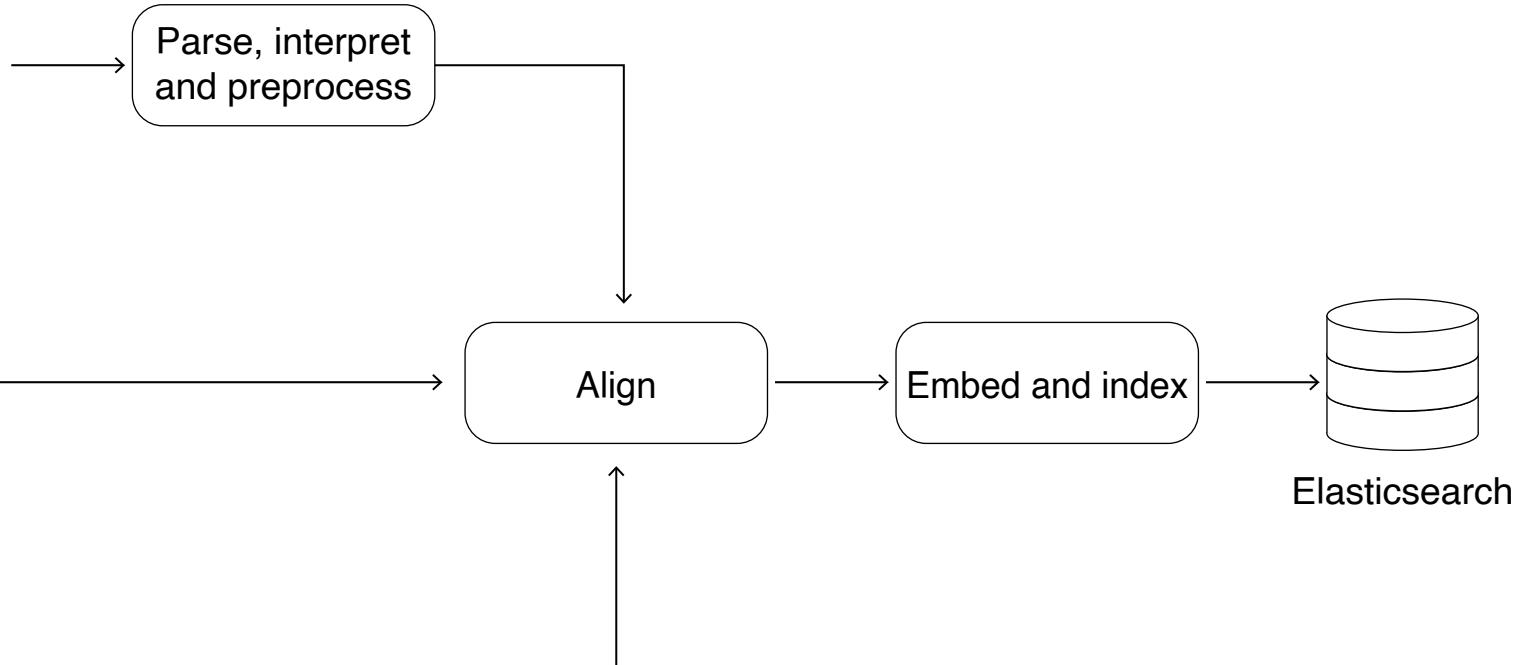
Latex sources

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

Inline comments



Lecture slides



Elasticsearch



— How do neural networks work?...



— How are they trained?



— How does backpropagation know which weight contributed most to the error?

## Query Representation

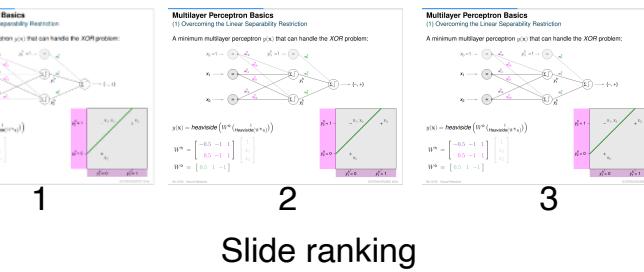
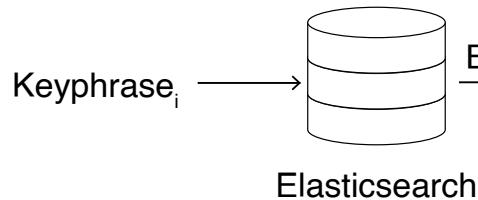
### 1a) Keyphrases:

- ❑ Extracted with KeyBERT
- ❑ {backpropagation, weight error, backpropagation weight error}

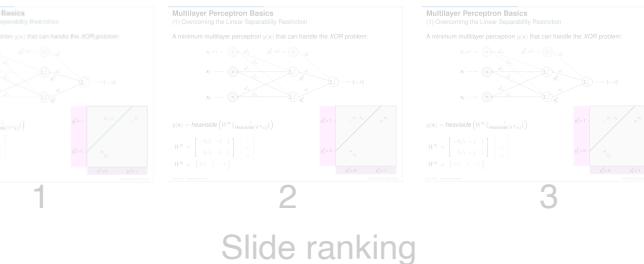
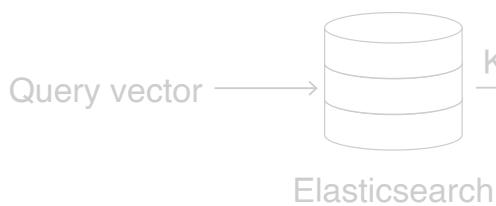
### 1b) Dense vector:

- ❑ SBERT embeddings
- ❑  $(-0.23, 0.56, 0.31, \dots)^T$

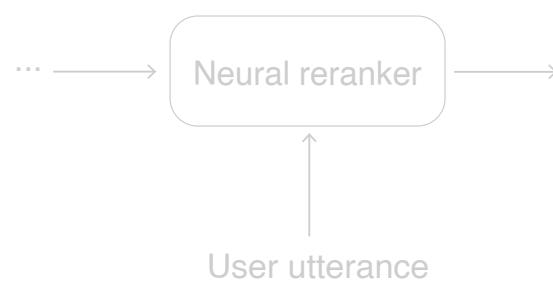
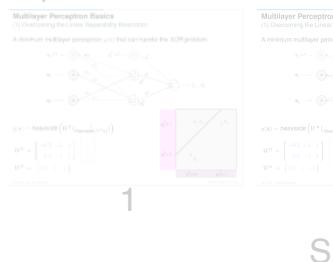
1a)



1b)



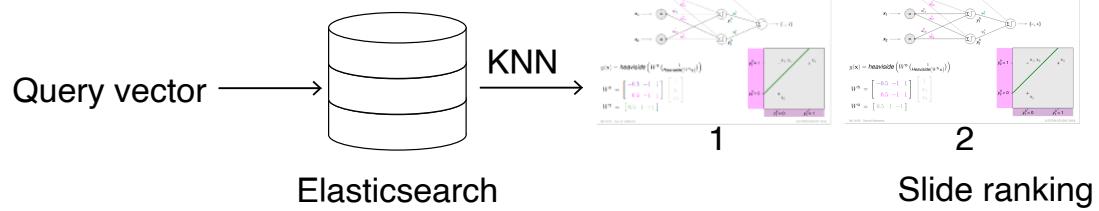
2)



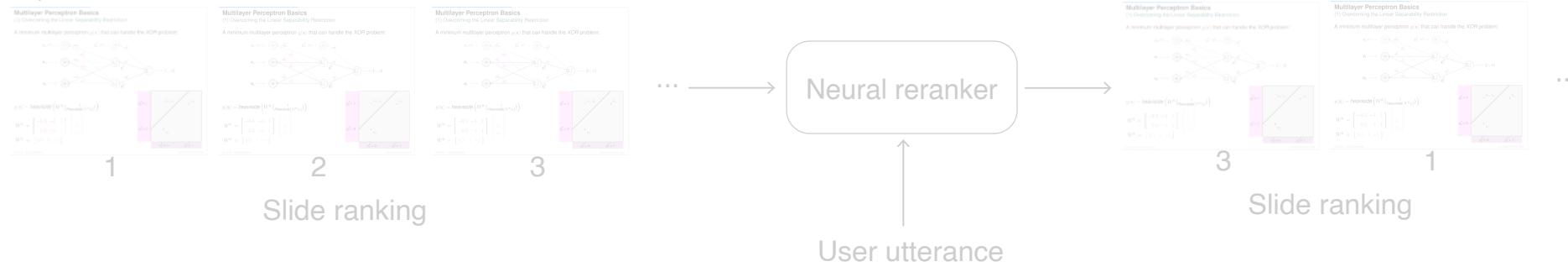
1a)



1b)

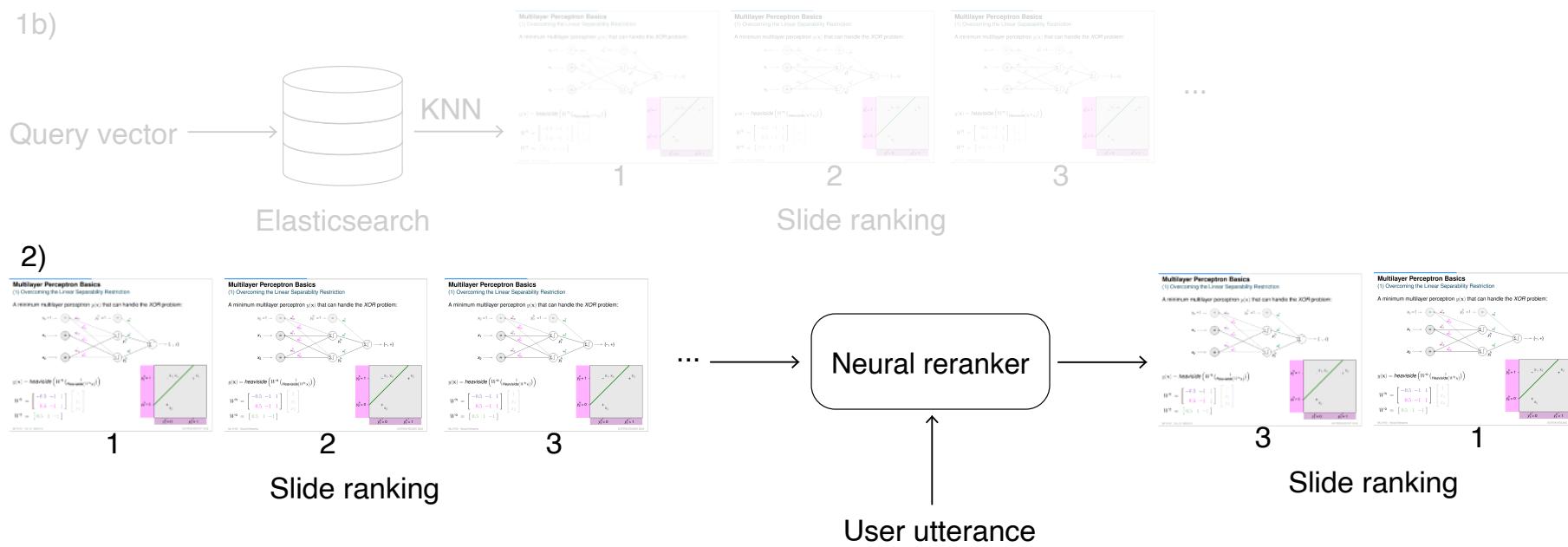


2)



1a)

1b)



You are a friendly teaching...

:

Use the following information to...

— Sys



— How do neural networks work?



Neural networks are used...



— How are they trained?



They are trained...



— How does backpropagation know which weight contributed most to the error?

Instruction-tuned LLM:

- ❑ Meta Llama 3
- ❑ 8 billion parameters
- ❑ Quantized to 6-bit floats

## 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

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

## Features:

- Embedded lecture slides
- References to online slides
- MathML equation rendering
- Focus+Context of responses and slides
- Time aligned speech-to-text with animated tokens
- Possibility to provide binary feedback for response

The screenshot displays the Infobot interface. On the left, a 'Sync Mode References' panel shows a box for 'Perceptron Learning' with a history of events from 1943 to 1965. On the right, a 'InfoBot' panel shows a conversation with a user named 'user' and the bot. The user asks 'Hi' at 11:25, and the bot responds with a welcome message and a thumbs-up icon at 11:25. The user then asks 'Can you Explain about The Perceptron of Rosenblatt?' at 11:30, and the bot provides a detailed explanation of the Perceptron model, including its mathematical formula:  $y(\mathbf{x}) = \begin{cases} 1, & \sum_{j=1}^p w_j x_j \geq \theta \\ 0, & \sum_{j=1}^p w_j x_j < \theta \end{cases}$ . The bot also mentions that this model was developed by Frank Rosenblatt in 1958. The interface includes a text input field, a 'Send' button, a 'Feedback' button, and a dropdown for 'Ryan (UK Male)'.

## 4) Experiments and Results

# ML Questions Dataset

(training split)

Manually crafted dataset of 100 questions and generated answers with binary feedback:



— What is a classification problem?



— What is a classification problem?

# ML Questions Dataset

(training split)

Manually crafted dataset of 100 questions and generated answers with binary feedback:



— What is a classification problem?

A classification problem means to assign a category or label to an input based on its features...



— What is a classification problem?

A classification problem is when a student in a school keeps switching classes and ...



# ML Questions Dataset

(training split)

Manually crafted dataset of 100 questions and generated answers with binary feedback:



— What is a classification problem?

A classification problem means to assign a category or label to an input based on its features...



— What is a classification problem?

A classification problem is when a student in a school keeps switching classes and ...



# ML Questions Dataset

(training split)

Manually crafted dataset of 100 questions and generated answers with binary feedback:



— What is a classification problem?

A classification problem means to assign a category or label to an input based on its features...



— What is a classification problem?

A classification problem is when a student in a school keeps switching classes and ...



Generator is trained on human feedback with Kahneman-Tversky Optimization (KTO).

# ML Questions Dataset (test split)

Manually crafted dataset of 101 question-answer pairs and relevant slides:

**Question:** “*What is overfitting?*”

Relevant Slides:

## Overfitting

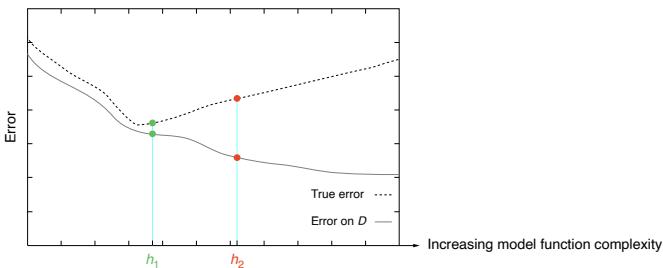
### Definition 9 (Overfitting)

Let  $D$  be a multiset of examples and let  $H$  be a hypothesis space. The hypothesis  $h_2 \in H$  is considered to overfit  $D$  if an  $h_1 \in H$  with the following property exists:

$$\text{Err}(h_2, D) < \text{Err}(h_1, D) \quad \text{and} \quad \text{Err}^*(h_1) < \text{Err}^*(h_2),$$

where  $\text{Err}^*(h)$  denotes the true misclassification rate of  $h$ , while  $\text{Err}(h, D)$  denotes the error of  $h$  on  $D$ .

[see continuation]



## Overfitting

### Mitigation Strategies

How to detect overfitting:

- Visual inspection

Apply projection or embedding for dimensionalities  $p > 3$ .

- Validation

Given a test set, the difference  $\text{Err}(y(), D_{\text{test}}) - \text{Err}(y(), D_{\text{tr}})$  is too large.

How to tackle overfitting:

- Increase the quantity and / or the quality of the training data  $D$ .

Quantity: More data averages out noise.

Quality: Omitting “poor examples” allows a better fit, but is problematic though.

- Early stopping of a model optimization / refinement process.

Criterion:  $\text{Err}(y(), D_{\text{val}}) - \text{Err}(y(), D_{\text{tr}})$  increases with increasing model function complexity.

- Regularization: Increase model bias by constraining the hypothesis space.

(1) Model function: Consider functions of lower complexity / VC dimension. [\[Wikipedia\]](#)

(2) Hypothesis  $w$ : Bound the absolute values of the weights in  $w$  of a model function.

Ground-truth Response: “*Overfitting occurs when a hypothesis  $h_2 \in H$  fits the training data. . .*”

# Experiments

## 1) Retrieval.

Cranfield-style IR experiment.

## 2) End-to-End.

Evaluate the full system with the Ragas framework.

# Experiments

## 1) Retrieval.

Cranfield-style IR experiment.

## 2) End-to-End.

Evaluate the full system 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?

# Results (retrieval)

Choice of ideal  $k$  for top- $k$  retrieval:

$$\operatorname{argmax}_k F_1^{\text{BM15}}((d_1, \dots, d_k)) = 6$$

$$\operatorname{argmax}_k F_1^{\text{KNN}}((d_1, \dots, d_k)) = 3$$

Retrieval pipeline	Precision@3	Recall@3	$F_1$
KNN + Reranking	0.40	0.48	0.44
KNN	0.38	0.46	0.42
Keyphrases + BM15 + Reranking	0.33	0.35	0.34
Keyphrases + BM15	0.31	0.33	0.32

# Results (end-to-end)

<b>Pipeline</b>	<b>Faithfulness</b>	<b>Correctness</b>	<b>Relevancy</b>
KNN + KTO	0.76	0.57	0.94
Keyphrases + BM15 + Reranking + KTO	0.73	0.55	0.91
KNN + Reranking + KTO	0.72	0.57	0.97
KNN + Reranking	0.71	0.59	0.97
:			
Keyphrases + BM15 + Reranking	0.68	0.56	0.96
Keyphrases + BM15	0.63	0.53	0.92



infobot.webis.de

## 5) Live Demo

## 6) Next Steps

# Next Steps

User study with computer science students:

- ❑ Does Infobot have positive impacts on knowledge gain?
- ❑ Does multi-modal representation influence memory retention?

Open technical challenges:

- ❑ Verbalization of mathematical equations.
- ❑ Coreference resolution in multi-turn dialog.
- ❑ Figure explanations and animations.
- ❑ Customization for teachers for their own courses.

# Next Steps

User study with computer science students:

- ❑ Does Infobot have positive impacts on knowledge gain?
- ❑ Does multi-modal representation influence memory retention?

Open technical challenges:

- ❑ Verbalization of mathematical equations.
- ❑ Coreference resolution in multi-turn dialog.
- ❑ Figure explanations and animations.
- ❑ Customization for teachers for their own courses.

**Thank you for your attention!**

© Marcel Gohsen 2025