

Facilitator Guide

Hands-On Methods for Teaching about Generative AI

Guide and material to facilitate the activities from the paper

Date	October 2026
Authors	Dr. Christophe Stammet, University of Luxembourg
Contact	christophe.stammet@uni.lu
License	This guide is licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Welcome and overview

1.1 About this guide

This serves as an overview to use the activities presented in the paper “Hands-On Methods for Teaching about Generative AI” to teach in school settings or in teacher trainings about fundamental concepts in the topic of generative artificial intelligence (AI) to an audience not requiring any prior technical knowledge.

1.2 The activities and their key ideas

Each activity is suitable for participants aged 15 upwards. A fully fledged workshop setting about generative AI can include all activities, they are also suited for separate use, depending on the contents you want to teach. This is an overview of the educational goals of each activity:

Key idea	Activity
Activity 1: How generative AI represents language	Language in generative AI systems is represented as points in a numerical embedding space, enabling models to capture similarities in meaning between words or parts of words.
Activity 2: How generative AI generates text	The primary training objective of text-based generative AI is to produce text in a statistically plausible way rather than to provide factually correct information, which can lead to convincing but erroneous content.
Activity 3: How generative AI represents bias	Generative AI systems inherit and can amplify social biases present in their training data, resulting in stereotypical representations of people and roles in generated content.

1.5 Activity organizational overview

Activity	Group size	Duration	Computer/tablet needed
1 Representing language	Groups of two or three	40'	No, Unplugged
2 Generating text	Individual	30'	Yes
3 Representing bias	Groups of two to four	40'	Yes

2. Activity 1: How generative AI represents language

2.1 At a glance

Key idea	Language in modern AI systems is represented as points in a numerical embedding space, enabling models to capture similarities in meaning between words or parts of words.
Learning objectives	Understand the concept of a list of numbers used to encode the meaning of a specific word. Realise the difficulty of attributing one specific value of meaning to each coordinate.
Group size	Groups of two or three participants
Duration	40'

2.2 Detailed run of the activity for facilitators

Concepts that are explained in this activity:

Tokenization, Embeddings, LLM training process.

The main goal of a Large Language Model being the capability to use and understand humanlike language, this activity introduces the first steps of creating an LLM. As any AI system requires numerical values as input, we need to transform text into numbers.

Participatory Question 1: How do computers transform text into numbers?

Usually, the first answer is "into binary" or "character by character". That is correct, but the goal here is to encode the meaning of the text, thus transforming individual characters do not carry enough information.

Participatory Question 2: Then what parts of the sentence could be transformed?

The reply to wait for here is "full words". This is correct, but after experiments, researchers found that an even better approach is transforming the text into tokens, which are a collection of common words, but also parts of words, symbols, etc., allowing an LLM to also understand other languages and less common words without requiring a massive database of words.

Participatory Question 3: How many tokens did the dataset for ChatGPT (2022) contain?

This questions gets a lot of different guesses, from hundreds of thousands into the millions and more. For GPT 3.5 (the LLM behind the first public release of ChatGPT), the number of tokens was approx. 50 000, a surprisingly small number for a model capable of conversion in English and also many other languages.

Key takeaway: A LLM learns the meanings of parts of words, so-called tokens. This is why e.g. earlier models had trouble counting the number of 'r's in the word 'strawberry' (this case went viral, highlighting limitations of LLM-assisted chatbots)

We now have text that is split into smaller tokens, we now need to train a LLM to understand the meaning of each token in its own way. The idea is that each token will be described by a series of numbers (i.e. a vector of numbers, depending of the level of the participants), with the idea that tokens with a similar meaning should also have a similar list of numbers.

The easiest way to illustrate these vectors is as a series of coordinates, on scales of meaning. To bring this concept closer, you can represent traditional coordinates as seen in mathematics or geography, but not as an "x- and y-axis", but axes of meaning (e.g. "where is the city from North (1) to South (-1)). Here you can introduce the limitation that every coordinate has to be between -1 and 1 (for advanced classes: this normalization process is done to guarantee that multiplying the values never escapes these bounds). This approach can easily add new columns to the coordinates by introducing new meaning (e.g. Population). A visual representation can be seen in Figure 1.

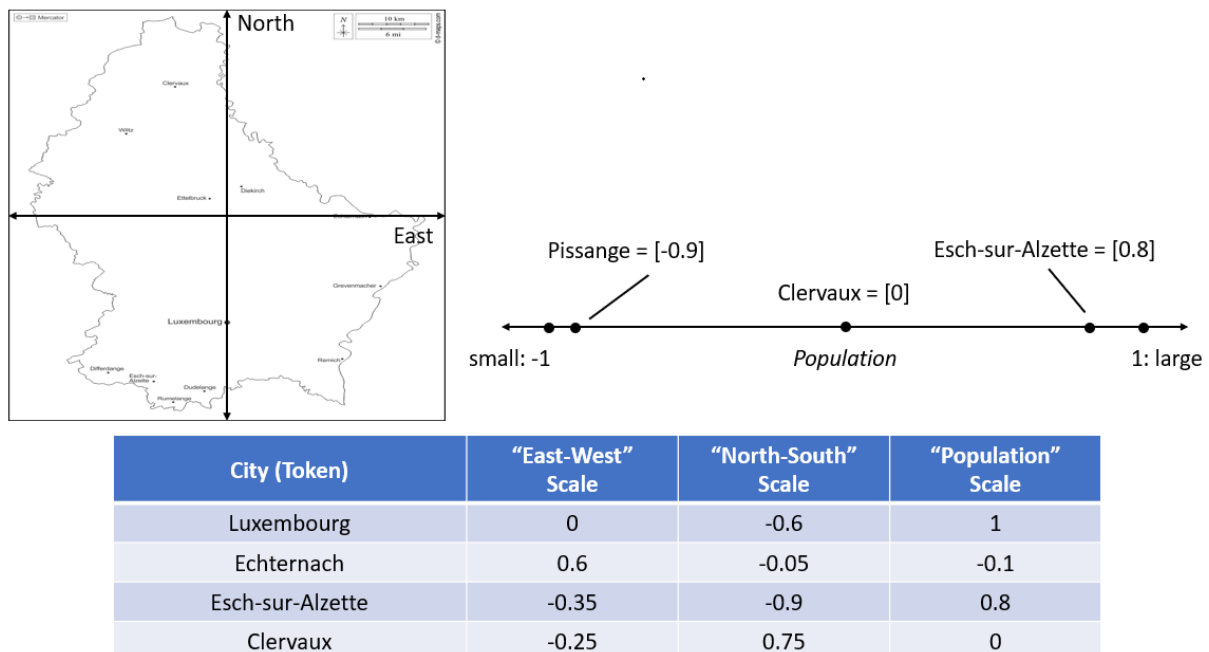


Figure 1: Visual representation of the embeddings as coordinates, also an additional "Population" scale is being introduced.

This leaning process of attributing coordinates based on a scale of meaning is now simulated by the participant in the practical activity. In small groups of 2-3 people, the participants get the hand-out as seen in Figure 2 (in English, can of course be adapted into other languages) and should work in their teams to fill out all the coordinates and the meaning of the 2nd and 3rd column.

Token	"Gender" Scale	"____" Scale	"____" Scale
Child		-0,8	
Man	-1		
Gal			-0,7
Baby			
Sir			1
Bro			-1
Girl		-0,5	
Granny		1	
Woman	1		

Figure 2: An example hand-out of the embedding exercise in English. The words can be adapted or translated in other languages.

Important information to share with the group:

- There is no right answer, they have to agree within a group
- They should start with the "gender" scale, then work on the next column
- Their interpretation of a word is what counts

After approx. 10 minutes, the teams should have finished the tables, and you can start discussing the results with the whole group.

2.3 Discussion

- **What are the names of your 2nd and 3rd scale?**

The second scale usually is called "age" scale by every group, there aren't really many other ways to interpret it (designed that way).

The third column however usually gets a lot broader results (e.g. "Formality", "Respect", "Niceness", "Status", ...). Let the groups share their answers and point out that they probably agree with every solution, even though they are slightly different.

- **Are the results of all the groups identical? If not, why?**

Usually not. It depends on the interpretation of the words and scales by each individual group.

- **What difficulties did you have filling out the table?**

It is difficult to fix one specific value to a word.

Some words have multiple meaning, which would greatly change their "coordinate" for certain meanings.

3. Activity 2: How generative AI generates text

3.1 At a glance

Key idea	The primary training objective of text-based generative AI is to produce text in a statistically plausible way rather than to provide factually correct information, which can lead to convincing but erroneous content.
Learning objectives	Analyse the different answers produced by a chatbot in terms of language proficiency and factual correctness.
Group size	Individual work
Duration	30'

3.2 Detailed run of the activity for facilitators

Concepts that are explained in this activity:

LLM text generation, Training, Hallucinations.

The way a Large Language Model writes text is the result of a statistical choice of the most probable next token based on the input prompt. An AI-Chatbot program (such as ChatGPT) uses such an LLM to sequentially generate an answer to the prompt.

Participatory Question 1: Is ChatGPT an AI? Yes or No?

This often gets the reaction of being a weird question with an obvious “Yes, of course”. However, one of the goals here is to understand that a Chatbot program such as ChatGPT merely uses an AI (the LLM-model it is using) to be able to provide a continuous text. As seen in Figure 3, the actual AI model only produces one token at a time, being repeatedly prompted by the program itself.

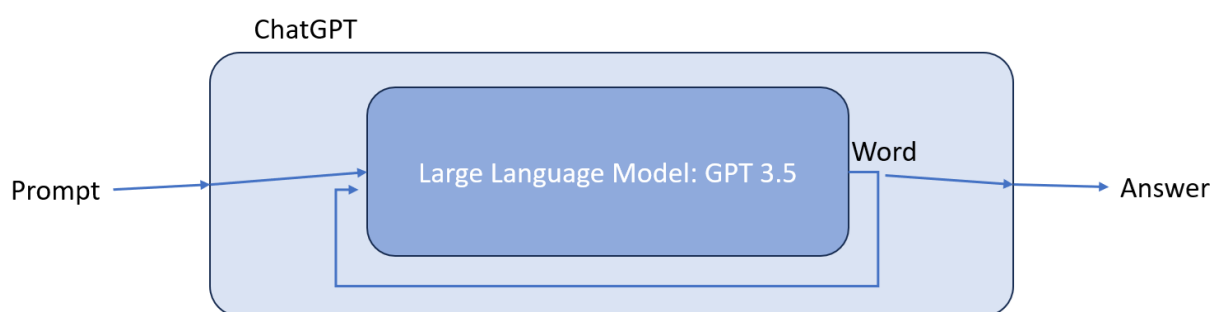


Figure 3: Simple overview of ChatGPT functionality in 2022. The LLM generates a single word, which is then added to the original prompt to provide the next output from the LLM.

The training process that leads to

This then leads into the free, experiment phase of this activity, interacting in a controlled way with a number of LLM models to discuss certain drawbacks.

Two possible platforms to use here are the following:

- Duck.ai: This platform lets you choose a number of LLMs to have a conversation with, no log-in needed.
- Arena.ai: This platform has access to some frontier models, and allows visualization side-by-side of two different models. Log-in required.

The best model to start this activity with is **GPT OSS 120B**, as this is a model with very advanced language capabilities, but lacks guardrails to remove content with factual uncertainty and is thus very prone to hallucinations. The following questions should then be asked by each participant to the LLM, once with and without internet access. After each question, the facilitator should discuss the main takeaways of each of the questions.

Write a biography of [your full name] from [your country of origin].

Sometimes requiring multiple prompting, this question usually leads to a well-written biography with completely fictitious content.

Give me all the details about this topic: [a topic you know well].

Here the results strongly depend on two factors: 1) does the LLM have Internet Search enabled and 2) is the topic known enough to be prevalent in the training data of the model.

Can you translate the following text into [Another language]: [Short copied text].

Many people are aware, but the power of Large **Language** Models lies in understanding language, and they are thus strong at translating or paraphrasing tasks, although the quality decreases for languages that have less content in the internet (e.g. Luxembourgish).

Key takeaway for participants:

Text generated by an LLM may produce factually incorrect information, as their main functionality lies in understanding and using human-like language, not guaranteeing factually correct content.

4. Activity 3: How generative AI represents bias

4.1 At a glance

Key idea	Generative AI systems inherit and can amplify social biases present in their training data, resulting in stereotypical representations of people and roles in generated content.
Learning objectives	Understanding how bias in training data can be amplified by generative AI.
Group size	Groups of two to four participants
Duration	40'

4.2 Detailed run of the activity for facilitators

Concepts that are explained in this activity:

Gender Bias, Text to image generation.

Text to image generation models have also emerged alongside and with LLM technology, as they share a similar functionality in understanding text as tokens and generating desired content from a prompt, images in this case.

The goal of this activity is to showcase how we can analyse the output of generative AI with respect to the eventual biases which are present in the training data. A simple first activity to showcase such biased behaviour on simpler AI models is to have a look at handwritten digit recognition. An excerpt of the training data can be seen in Figure 4.



Data & Labels

Figure 4: Excerpt from the MNIST dataset used to train handwritten digit recognition models.

When using such a tool (some online tools are available, downloadable models can also be found online), a perfect example to showcase the bias of these written digits is when writing the number 1 in the European way of writing digits, as can be seen in Figure 5.

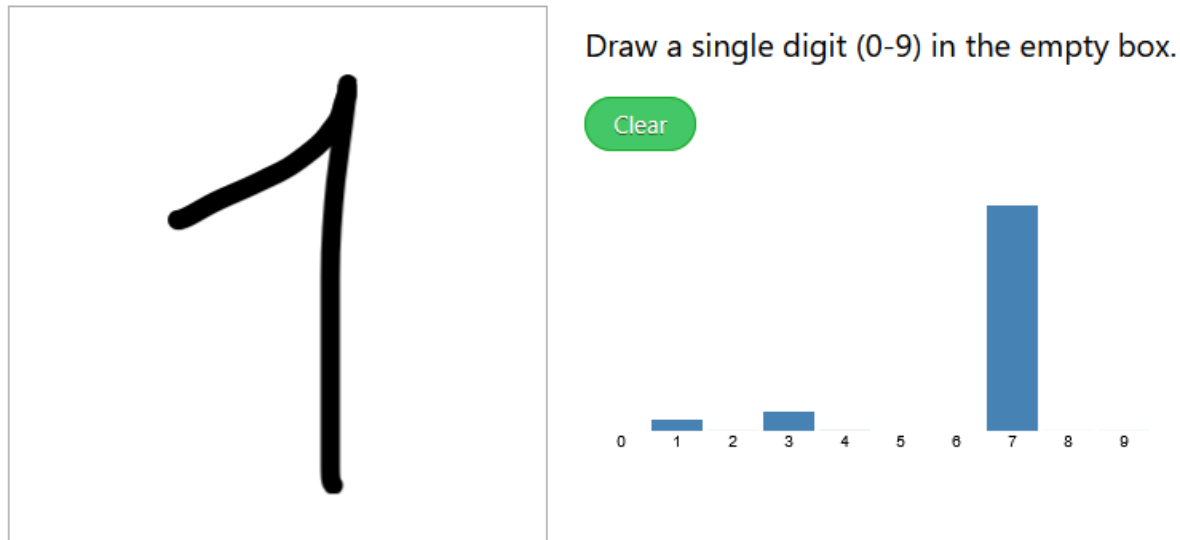


Figure 5: A handwritten number 1 and the result of the AI classifier recognizing mainly a 7.

When analysing the training data, it is clear that the handwritten digits are biased towards the e.g. American way of writing digits, where the 1 is mainly just a vertical dash and the 7 does not contain a dash through the middle.

For the main activity, we will look at the gender bias present in the data when looking at gender-neutral words (nurse, computer scientist, ...) and is split into the following two questions the participants need to answer with experiments. The handout on the last page of this chapter combines both experiments onto one page to be distributed to each group.

1. Is the training data of text-to-image models biased?

Most text-to-image generators have been trained on publicly available images from the internet to be able to combine information present in the text descriptions of the image with the actual image (e.g. the dataset LAION 5B, containing 5 billion image-text pairs of images).

A way to illustrate the images in training data is to look at image search engines (e.g. Google, Ecosia, DuckDuckGo, ...) and label each image based on their visual inference of the person shown and tally the results on the handout, as seen in Figure 6..

	Internet-Sich		total
weiblech	Google  Bing 	Ecosia  DuckDuckGo 	45
männlech	Google  Bing 	Ecosia  DuckDuckGo 	25
Net secher / keng / non-binär	Google  Bing 	Ecosia  DuckDuckGo 	5

Figure 6: Example result (in Luxembourgish) of the results of the first experiment. Gender distribution based on visual inference for the word "nurse".

2. How does this affect the bias of AI-generated images?

To answer the second question, all we need to do is generate images using an AI tool and repeat the counting process from the first step. Any image generation tool that is trained on publicly available data works for this (<https://magicstudio.com/ai-art-generator/> is free to use and login-free).

After having the participants gather the data for AI-generated imagery, the facilitator can collect the data from both steps of the activity to visualize the results, as seen in Figure 7.

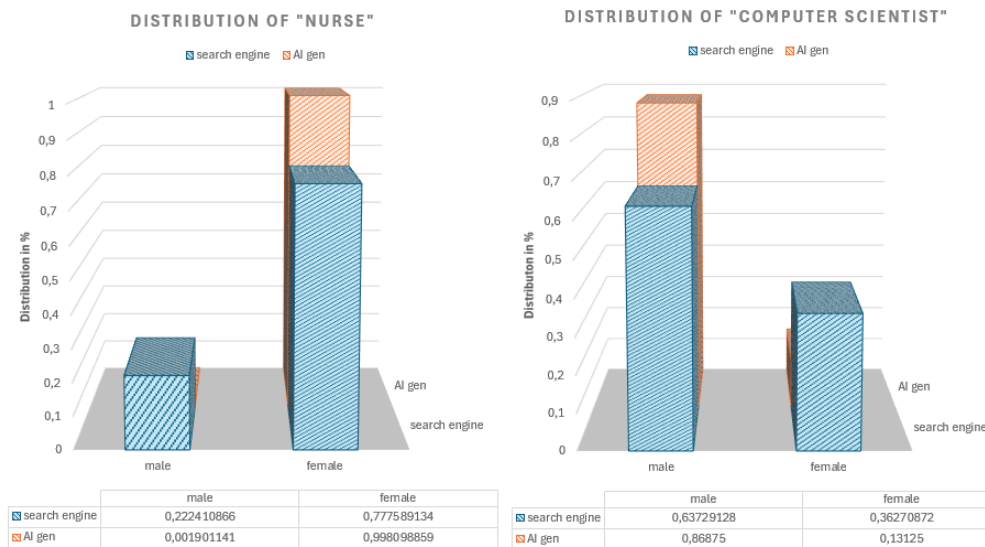


Figure 7: Results from the image search (in blue) and the AI-generated images (in orange) for the search words "nurse" and "computer scientist".

When discussing these results, it is important to emphasize that such biased behaviour also occurs in AI-generated text, but is harder to visualize than the image generation.

Alternative keywords:

If you want to avoid the discussions that stem from discussing gender only based on appearance or discounting non-binary people, you can also use this experiment with a large variety of other biases present in the data. Some examples:

- "a glass of wine" (red wine being overrepresented)
- "a pizza" (peperoni pizza being overrepresented)
- "a hand writing text" (right hands primarily)

Key takeaway for participants:

The result of AI-generated content will always be directly dependent on the biases present in the training data.

4.3 Printable handout of the activity

	Internet search	Total count	Total %%
female			
male			
Not sure / none / non-binary			
Step 1: Go to one or more search engines (e.g. google.com, duckduckgo.com, bing.com, ecosia.com) and use Image Search to look for images for your word (in the top left corner) In the boxes above, make a tick for each person's gender.			

	AI-generated images	Total count	Total %%
female			
male			
Not sure / none / non-binary			
Step 2: In your browser, use the magicstudio.com automatic image generator. As prompt, use only your word (in the top left corner)			

Step 3: Calculate the distribution of genders for the total of images and compare the distribution in the internet search compared to the AI generated images. Is there a bias in the training data? Is there a bias in the AI generated images? If so, is it higher or lower than in the training data?

5. Feedback and Contact

5.1 Feedback

If you facilitate a session with one or multiple activities that were presented in this document, we would greatly appreciate getting immediate feedback from you.

A short text regarding target audience, number of participants, age of participants and which activities you did with the group.

Please let us know of any facilitated sessions, their outcomes and reaction by the participants directly under the contact seen below

5.2 Contact

If you require any further information or additional context regarding the activities presented in this document, feel free to contact the author directly:

Dr. Christophe Stammet
Education Programme Specialist

christophe.stammet@uni.lu

(+352) 46 66 44 9983



Université du Luxembourg

Scienseens Lab

7, avenue des Hauts-Fourneaux
L-4362 Esch-sur-Alzette